UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

EXTENSÃO DO *BAG-OF-VISUAL-FEATURES* PARA INCORPORAR INFORMAÇÃO ESPACIAL NA DESCRIÇÃO DE CARACTERÍSTICAS DE ACORDO COM A PERCEPÇÃO VISUAL HUMANA

ROBSON DE CARVALHO SOARES

Uberlândia - Minas Gerais
2012

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

ROBSON DE CARVALHO SOARES

EXTENSÃO DO *BAG-OF-VISUAL-FEATURES* PARA INCORPORAR INFORMAÇÃO ESPACIAL NA DESCRIÇÃO DE CARACTERÍSTICAS DE ACORDO COM A PERCEPÇÃO VISUAL HUMANA

Dissertação de Mestrado apresentada à Faculdade de Computação da Universidade Federal de Uberlândia, Minas Gerais, como parte dos requisitos exigidos para obtenção do título de Mestre em Ciência da Computação.

Área de concentração: Banco de Dados.

Orientadora:
Prof[a]. Dr[a]. Denise Guliato

Uberlândia, Minas Gerais
2012

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Os abaixo assinados, por meio deste, certificam que leram e recomendam para a Faculdade de Computação a aceitação da dissertação intitulada "**Extensão do *bag-of-visual-features* para incorporar informação espacial na descrição de características de acordo com a percepção visual humana**" por **Robson de Carvalho Soares** como parte dos requisitos exigidos para a obtenção do título de **Mestre em Ciência da Computação**.

Uberlândia, 29 de Fevereiro de 2012

Orientadora: _____

Profª. Drª. Denise Guliato
Universidade Federal de Uberlândia

Banca Examinadora:

_____

Profª. Drª. Celia Aparecida Zorzo Barcelos
Universidade Federal de Uberlândia

_____

Prof. Dr. Ilmério Reis da Silva
Universidade Federal de Uberlândia

_____

Prof. Dr. Ricardo da Silva Torres
Universidade de Campinas

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Data: Fevereiro de 2012

Autor:	**Robson de Carvalho Soares**
Título:	**Extensão do *bag-of-visual-features* para incorporar informação espacial na descrição de características de acordo com a percepção visual humana**
Faculdade:	**Faculdade de Computação**
Grau:	**Mestrado**

_____

Autor

Dedicatória

À minha amada esposa Natassia Catita Silva de Carvalho e à minha querida família Joaquim Soares Sobrinho, Aparecida Mafra de Carvalho Soares, Pollyanna Mafra Soares e Mayara Mafra Soares.

Agradecimentos

Gostaria de agradecer...

Principalmente a Deus, por não só me proporcionar mais esta oportunidade mas também por me dar a capacidade e a força necessária para enfrentá-la.

À minha amada esposa e companheira Natassia Catita Silva de Carvalho, que além de todo amor e carinho, SEMPRE me apoiou e me incentivou mesmo que isso pudesse causar a minha ausência em momentos tão importantes.

Aos meus queridos e exemplares pais Joaquim Soares Sobrinho e Aparecida Mafra de Carvalho Soares que de certa forma foram os responsáveis pela minha matrícula no mestrado. Sem a insistência, paciência, confiança, carinho e amor deles, eu não teria conseguido.

Às minhas queridas irmãs Mayara Mafra Soares e Pollyanna Mafra Soares, pelo amor, carinho e paciência.

Em especial à minha orientadora Denise Guliato, pela amizade, confiança, apoio, paciência e orientação nesta pesquisa.

Ao professor Ilmério Reis da Silva, pela co-orientação dessa pesquisa.

Aos meus amigos do laboratório Ernane, Ricardo, Juliano, Jean, Elaine, Allan, Walter, Vinícius, Dali e Ivo pela amizade e apoio no que fosse preciso.

À amizade e companheirismo de Tarcísio A. de Magalhães Júnior e Lígia Maria Soares Passos desde a graduação até os dias de hoje. Vale lembrar as rotinas árduas de estudo que varavam as madrugadas.

À empresa SOFTBOX e todos os meus colegas de trabalho que SEMPRE me apoiaram nessa empreitada. Em especial aos meus grandes amigos Daniel Vincenzi e Orseni Ferreira Campos pela confiança, apoio e paciência.

Aos meus grandes amigos e irmãos da faculdade e da vida Tarsa, Dudu, Guilermo, Franz, Baiano, Jarraum, Sabaum, Pedro, Itaú, Paum, Pelanquinha, Tora, Minero, Carlim, Dino, Jardel, Digão, Danilão, Lígia, Andressa, Luciana, Natália, Loló, Raquel e Lulu por não concordarem mas entenderem o porquê das minha ausências.

Resumo

Bag-of-visual-features é uma abordagem bem conhecida pela classificação de padrões e recuperação de imagens por conteúdo. Toda a imagem é descrita por um histograma normalizado de frequências de palavras visuais. Neste caso, objetos de interesse em grandes áreas e/ou diferentes cenas de fundo, podem não ser corretamente reconhecidos. Neste trabalho, propomos extrair as partes relevantes da imagem (*foreground*) de acordo com a percepção visual humana utilizando a abordagem de extração de mapas de saliências. Em nossa proposta, o histograma é construído ponderando as palavras visuais de acordo com o mapa de saliência em duas abordagens, *fuzzy* e binária, incorporando uma distinção entre informação de fundo (*background*) e primeiro plano (*foreground*). Nosso método foi testado usando bases de dados de imagens com diferentes condições de iluminação, cores, posicionamento, escala, e mudanças de fundo. A análise dos resultados demonstra que a nossa proposta apresenta melhorias significativas sobre outras abordagens testadas.

Palavras chave: *content-based image retrieval*, *bag-of-visual-features*, dicionário de palavras, versões de imagem, sift, histogramas

Abstract

Bag-of-visual-features is a well-known approach for pattern classification and content-based image retrieval. The entire image is described by a normalized histogram of frequencies of visual words. In this case, objects of interest on large background area or on different background scenes, may not be correctly recognized. In this work we propose to extract the relevant scenes of the image (foreground) according to the human being visual perception using the saliency map approach. In our proposal, the histogram is built on weight visual words according to the saliency map on two approaches, fuzzy and binary, to incorporate a distinguish between information of background and foreground. Our method was tested on databases that contain images with different conditions of illumination, color, rigid and scale transformations, and changes of the background. The analysis of the results demonstrates that our proposal presents significant improvements over other tested approaches.

Keywords: content-based image retrieval, bag-of-visual-features, codebook, image's version, sift, histograms

Sumário

Lista de Figuras

Lista de Abreviaturas e Siglas

SGBD	Sistema Gerenciador de Banco de Dados
CBIR	*Content-Based Image Retrieval*
TBIR	*Text-Based Image Retrieval*
RQ	*Range Query*
KNNQ	*K-Nearest Neighbor Query*
WTA	*Winner-Take-All*
AveP	*Average Precision*
MAP	*Mean Average Precision*
DCG	*Discount Cumulative Gain*
NDCG	*Normalized Discount Cumulative Gain*
GBVS	*Graph-Based Visual Saliency*
SIFT	*Scale Invariant Feature Transform*
SCD	*Scalable Color Descriptor*
CLD	*Color Layout Descriptor*
EHD	*Edge Histogram Descriptor*
CEDD	*Color and Edge Directivity Descriptor*
FCTH	*Fuzzy Color and Texture Histogram*
BIC	*Border-Interior Pixel Classification*
LCPC	*Local Color Pixel Classification*
DCT	*Discrete Cosine Transform*
CLF	*Coordinate Logic Filters*
DoG	*Difference of Gaussian*
MP	*Mapa de Saliência*

Capítulo 1

Introdução

O avanço de tecnologias tem colaborado com a evolução de dispositivos cada vez mais robustos para a aquisição e armazenamento de imagens, gerando um aumento muito grande e rápido do volume de informações multimídia em bases de dados em áreas como a medicina, geografia, engenharia entre outras. Dessa maneira, nossa capacidade de capturar e armazenar dados ultrapassou largamente a nossa capacidade de processá-los e utilizá-los eficientemente, podendo acontecer de informações armazenadas nunca mais serem acessadas [Fayyad e Uthurusamy 2002]. Além disso, o acesso eficiente a imagens em grandes bases de dados é ainda mais problemático quando estas bases não estão organizadas [Rui et al. 1997], como é o caso das bases disponíveis na Web. Neste cenário, surge a necessidade da criação de técnicas cada vez mais eficazes para o armazenamento e recuperação de imagens.

Os primeiros trabalhos relacionados ao processo de recuperação de imagens foram apresentados no final dos anos 70 e início dos anos 80. Trabalhos como os de [Chang e Fu 1980] e [Chang e Kunil 1981] associam informações textuais às imagens para uma posterior consulta baseada em texto conforme os sistemas de gerenciamento de base de dados tradicionais. Essa técnica, conhecida como recuperação de imagens baseada em texto (TBIR - *Text-Based Image Retrieval*), pode utilizar as seguintes informações textuais associadas às imagens: descrições, legendas, palavras-chaves ou meta-dados como tipo de exame, corte, número de *pixels*, bits por *pixel*, equipamento de geração, data e hora da criação, posicionamento global, caminho no sistema de arquivos, configurações de brilho e contraste, entre outros. As dificuldades dessa abordagem são que as informações textuais têm que ser associadas a todas as imagens, por uma ou mais pessoas, o que torna um trabalho muito demorado e cansativo. Além disso, diferentes interpretações podem ser dadas para a mesma imagem por pessoas diferentes devido à subjetividade humana.

Assim, como proposta para evitar essas dificuldades eliminando principalmente o processo mecânico de anotações de imagens, surgem as técnicas de Recuperação de Imagens Baseada em Conteúdo (CBIR - *Content-Based Image Retrieval*) [Kato 1992] que utilizam características de cor, textura e forma extraídas automaticamente das imagens e repre-

sentadas por valores reais [Traina et al. 2003], [Nascimento et al. 2003] [Alto et al. 2005]. A cada imagem é associado um vetor de características. A distância entre dois vetores de características indica o grau de similaridade entre as respectivas imagens. Esta abordagem é utilizada na maioria dos sistemas CBIR encontrados na literatura [Carson et al. 1999], [Niblack et al. 1993] e [Smith e fu Chang 1996].

Recentemente, uma nova técnica conhecida como *bag-of-visual-features* ou ainda *bag-of-visual-words* [Sivic e Zisserman 2003], [Csurka et al. 2004] vem sendo utilizada na recuperação de imagens por conteúdo. Essa abordagem tem se mostrado computacionalmente eficiente e tem sido bem sucedida para categorização de objetos e cenas. O método faz analogia a técnica conhecida como *bag-of-words* [Dumais et al. 1998] e [Ribeiro-Neto et al. 1999] utilizada em recuperação de informações textuais, na qual cada documento é representado como um conjunto das palavras que ocorrem nesse documento.

Na abordagem *bag-of-visual-features* são criados dicionários de palavras visuais que são gerados pelo agrupamento dos vetores de características das imagens. Cada palavra visual representa, portanto, um padrão local específico compartilhado por todos os descritores de um dado agrupamento. Para cada imagem é gerado um histograma de frequência das palavras visuais que corresponde à frequência com que uma determinada palavra visual aparece na imagem. A partir desses histogramas é possível realizar a recuperação de imagens.

O dicionário de palavras visuais gerado pelo agrupamento dos vetores de características obtidos por algum descritor, como por exemplo, o SIFT [Lowe 2004] não leva em consideração a diferenciação entre o *background* e *foreground* da imagem. Um dos motivos é a dificuldade de identificar qual parte da imagem pertence ao *background* e qual pertence ao *foreground* de uma maneira fácil, rápida e automática. Alguns trabalhos na literatura têm explorado essa diferenciação em estudos da percepção visual humana como é o caso de [Moosmann et al. 2006], [Marszalek e Schmid 2006], [Huang et al. 2008], [Sato e Katto 2010] entre outros.

Encontramos uma resposta para esta diferenciação em estudos da percepção visual humana. Uma técnica denominada Mapas de Saliência [Itti et al. 1998] e [Harel et al. 2007], quando aplicada a uma imagem é capaz de identificar o que é relevante na imagem de acordo com a percepção humana. Portanto, neste trabalho são exploradas maneiras diferentes de se construir o histograma de palavras visuais considerando a percepção visual humana com o intuito de melhorar a qualidade de recuperação das informações.

1.1 Objetivos

Este trabalho explora a abordagem *bag-of-visual-features* para o problema da recuperação de imagens por conteúdo. Propomos duas estratégias diferentes para descrever imagens considerando a percepção visual humana, a qual nos permite realizar a distinção

entre o primeiro plano (*foreground*) e o fundo (*background*) da imagem via a extração dos mapas de saliência. As duas estratégias estendem a abordagem *bag-of-visual-features* modificando o processo de construção de histogramas de frequência de palavras visuais.

Na primeira estratégia, o histograma é construído considerando a distinção do *background* e do *foreground* de forma binária utilizando o mapa de saliência da imagem para identificar se uma determinada palavra visual pertence ao *background* ou *foreground* da mesma. Já na segunda estratégia, essa distinção ocorre de forma *fuzzy* e dessa maneira o mapa de saliência é utilizado para definir o grau de pertinência das palavras visuais da imagem em relação ao seu *background* e *foreground*.

Neste trabalho também foram criadas oito bases de imagens com um total 4720 imagens para que os métodos propostos fossem avaliados quanto a capacidade de invariância a mudança de cor, iluminação, escala, rotação, translação e fundo das imagens. Essas bases são consideradas bases de versões, uma vez que são geradas diversas versões de uma determinada imagem para compor a base.

1.2 Organização do Trabalho

A dissertação está organizada em seis capítulos. No Capítulo 2 são apresentados os conceitos teóricos utilizados neste trabalho; são detalhadas algumas técnicas como CBIR, *bag-of-visual-features* e extração de atenção visual. Uma revisão bibliográfica dos principais trabalhos encontrados na literatura relacionados à utilização da percepção visual humana no processo de recuperação de imagens é apresentada no Capítulo 3. As estratégias propostas neste trabalho são descritas no Capítulo 4. No Capítulo 5 são apresentados os experimentos realizados descrevendo as bases de dados utilizadas, a metodologia de avaliação e os resultados obtidos. Por fim, o Capítulo 6 relata as conclusões obtidas, as contribuições realizadas e os trabalhos futuros propostos.

Capítulo 2

Fundamentação Teórica

Este capítulo tem como objetivo apresentar os conceitos teóricos necessários para o desenvolvimento deste trabalho.

2.1 Recuperação de Imagens por Conteúdo

O processo de recuperação de imagens por conteúdo (CBIR - *Content-Based Image Retrieval*) [Kato 1992] tem como objetivo buscar imagens similares a uma imagem de consulta. O processo inicia-se pela extração de características das imagens da base. São utilizados os chamados descritores de imagem para extrair informações referentes ao conteúdo visual de cada imagem. O conteúdo a ser extraído pode ser referente a cor, a forma ou a textura da imagem. Assim, o conjunto de diversas características extraídas pelos descritores forma os chamados vetores de características. Estes vetores são comparados via funções de distância com o intuito de definir a similaridade das imagens da base em relação à imagem utilizada na consulta.

O fluxo de funcionamento da maioria dos sistemas CBIR é mostrado na Figura 2.1. Inicialmente, para todas as imagens inseridas na base de consulta, são gerados seus vetores de características armazenando-os na base de dados. O usuário fornece ao sistema uma imagem de referência da qual também são extraídas suas características visuais e formado seu vetor de características. Assim, o sistema realiza o cálculo da similaridade entre os vetores de características da imagem de referência e os vetores armazenados na base de dados, utilizando algum método de indexação para melhorar o desempenho da consulta. Alguns sistemas ainda utilizam o processo conhecido como *relevance feedback* com o intuito de gerar uma recuperação semanticamente mais significativa. Neste processo o usuário tem a possibilidade de refazer a consulta refinando o resultado pela seleção das imagens que mais foram relevantes, de acordo com o seu ponto de vista, no resultado obtido [Feng D. 2003].

A seguir são apresentados os conceitos dos principais componentes de um sistema CBIR como os métodos de extração de características visuais das imagens, medidas de

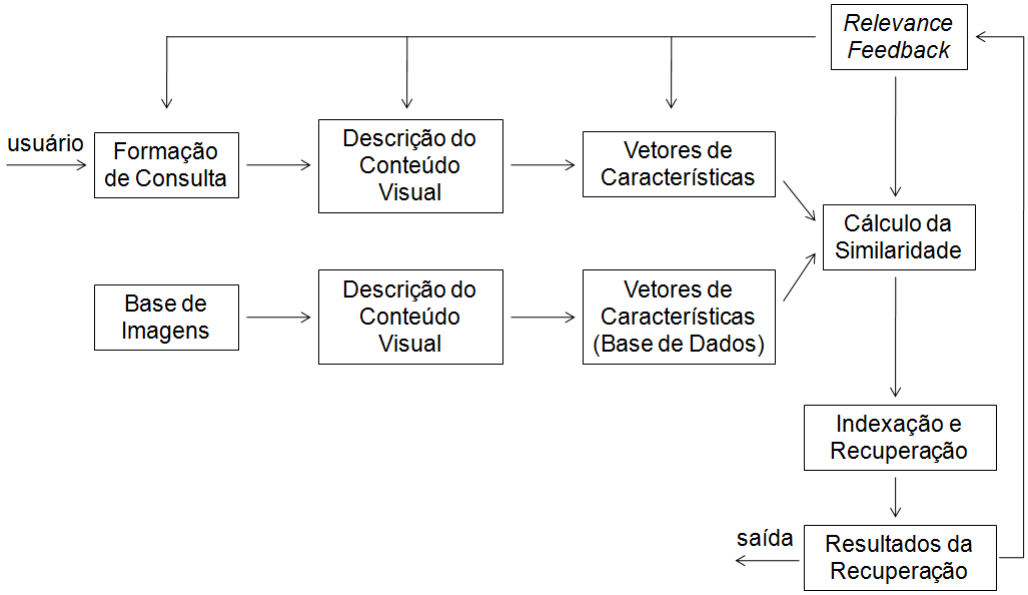

Figura 2.1: Fluxo de funcionamento de um sistema CBIR. Ilustração retirada do artigo [Feng D. 2003].

distâncias utilizadas para calcular a similaridade entre as imagens e operadores de similaridade responsáveis pelas consultas.

2.1.1 Extração de Características

Em um sistema CBIR o processo de extração de características de imagens corresponde à obtenção automática de informações visuais das imagens que possam ser manipuladas por um mecanismo de recuperação. Os descritores são responsáveis por extrair tais características as quais são organizadas em vetores de características. Os vetores são utilizados então na recuperação por conteúdo, ao invés das imagens propriamente ditas.

As principais características visuais de uma imagem utilizadas no processo de recuperação por conteúdo são: cor, forma e textura. A seguir são apresentadas técnicas de extração e representação dessas características.

Cor

A cor é uma das características mais utilizadas para representar uma imagem no processo de recuperação de imagens por conteúdo devido ao baixo custo computacional. Os modelos de representação de cores mais conhecidos são o RGB (*red, green, blue*), que mapeia diretamente as cores do componente de exibição da imagem e o HSI (*hue, saturation, intensity*) que mapeia mais fielmente um modelo de cores para a percepção humana [Gonzalez e Woods 2006]. A representação das cores de uma imagem em um vetor de características é mais comum a partir de histogramas. O histograma de cores para recuperação de imagens por conteúdo foi introduzido por [Swain e Ballard 1991].

Os histogramas representam a quantização dos espaços de cores pela contagem do número de *pixels* que cada cor quantizada possui na imagem. Eles são invariantes à translação e rotação das imagens, mas, não indicam a localização espacial dos *pixels* na imagem. Histogramas de níveis de cinza podem ser extraídos diretamente de imagens coloridas. A Figura 2.2 mostra um exemplo de um histograma de 256 níveis de cinza.

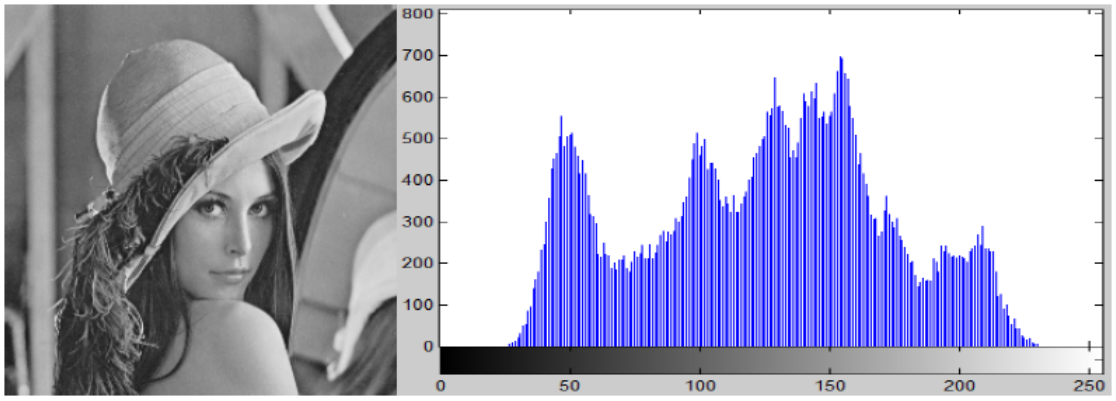

Figura 2.2: Exemplo de uma imagem e seu histograma quantizado em 256 níveis de cinza.

Embora essas técnicas ofereçam formas satisfatórias para caracterizar a informação de cor, elas sofrem pela falta de informação espacial do conteúdo visual de uma imagem, pois falham em distinguir imagens com a mesma cor, mas com uma distribuição diferente [Kimura et al. 2011]. Assim, percebemos que não se pode utilizar somente os histogramas para definir a similaridade das imagens, pois duas imagens completamente diferentes podem possuir histogramas similares.

Uma maneira de incluir informação espacial sobre o conteúdo visual de imagens é a utilização de descritores de cor locais. Descritores de cor locais podem ser classificados em dois grupos: abordagens de regiões e abordagens baseadas em partições. Abordagens de regiões dividem a imagem em regiões, que podem ser diferentes para cada imagem. Embora essas abordagens têm se mostrado efetivas [Pass et al. 1996], elas frequentemente usam algoritmos de segmentação automática da imagem que, em geral, possuem alto custo computacional e processos complexos de extração de características visuais, tornando a aplicação em grandes bancos de dados de imagens inviável [Kimura et al. 2011].

Por outro lado, as abordagens baseadas em partições incluem informação espacial de características visuais particionando a imagem em blocos de tamanho fixo e em seguida extraindo as características de cada bloco individualmente. Esta técnica é mais simples do que as abordagens de regiões, já que o mesmo esquema de particionamento é aplicado a todas as imagens [Kimura et al. 2011].

Seguem abaixo alguns exemplos de outras técnicas de extração de propriedades de cor da imagem, a maioria destes descritores são classificados como abordagens baseadas em partições:

- **SCD - *Scalable Color Descriptor*:** O SCD [Manjunath et al. 2001] é um histo-

grama de cor codificado pela transformada de Haar [Stollnitz et al. 1996] com uma quantização uniforme do espaço de cor HSV em 256 posições. Este descritor utiliza a distância L_1 (apresentada na subseção 2.1.2) como métrica de similaridade entre duas imagens.

- **CLD - *Color Layout Descriptor*:** O CLD [Manjunath et al. 2001] tem a funcionalidade de capturar a distribuição espacial de cor de qualquer região da imagem. Inicialmente, a imagem é dividida em 64 blocos (8x8) onde a cor dominante de cada bloco é extraída, geralmente, pela média de cor. Posteriormente, cada componente do espaço de cor YCrCb é transformado utilizando a transformação cosseno (DCT - *Discrete Cosine Transform*) [Ahmed et al. 1974] gerando três conjuntos de coeficientes. Finalmente, um peso é associado a cada coeficiente produzindo um vetor de características com informação das cores predominantes em cada bloco. Para calcular a similaridade entre duas imagens A e B, dois conjuntos de coeficientes são considerados, o primeiro conjunto pertence à imagem A e o segundo à imagem B, em que A = {DY, DCb, DCr} e B = {DY', DCb', DCr'}. Assim, a similaridade entre A e B pode ser obtida pela Equação 2.1:

$$\begin{aligned} D &= \sqrt{\sum_i w_{yi}(DY_i - DY_i')^2} \\ &+ \sqrt{\sum_i w_{bi}(DCb_i - DCb_i')^2} \\ &+ \sqrt{\sum_i w_{ri}(DCr_i - DCr_i')^2} \end{aligned} \tag{2.1}$$

em que (DY_i, DCb_i, DCr_i) representa o coeficiente DCT i do respectivo componente de cada cor. Os pesos são devidamente atribuídos, e um componente com menor freqüência recebe o maior peso. Figura 2.3 apresenta os valores de ponderação recomendados para cada coeficiente.

	0	1	2	3	4	5
Y	2	2	2	1	1	1
Cb	2	1	1			
Cr	4	2	2			

Figura 2.3: Pesos recomendados para os coeficientes do descritor CLD.

- **EHD - *Edge Histogram Descriptor*:** Este descritor [Manjunath et al. 2001] corresponde a um extrator de características de textura (descrito no próximo tó-

pico desta seção 2.1.1) da imagem. Ele é utilizado para compor as características extraídas pelo próximo descritor a ser apresentado, o CEDD. Para extrair as características, a imagem é dividida em 16 sub-imagens. Para cada sub-imagem é montado o histograma local de arestas. Essas arestas são agrupadas em 5 categorias: vertical, horizontal, 45^o diagonal, 135^o diagonal e sem nenhuma orientação específica. Assim, cada histograma possui 5 posições correspondendo a esses 5 tipos de arestas e como são 16 sub-imagens, o histograma final da imagem possui 80 posições.

A Figura 2.4 mostra como os histogramas locais de arestas são calculados. Inicialmente, cada uma das 16 sub-imagens são subdividas em blocos da imagem. O tamanho de cada bloco corresponde a uma escala do tamanho da imagem elevado a potência de 2. O número de blocos por sub-imagem é constante. Em seguida, cada bloco é dividido em blocos de 2x2, sendo que a média das intensidades dos *pixels* desses novos blocos formam a representação da partição. Dessa forma, cada partição é tratada como um pixel, e os operadores detectores de arestas são aplicados nessas partições. O EHD utiliza até cinco filtros detectores de arestas (Figura 2.5) nas orientações utilizadas. A distância L_1 (apresentada na subseção 2.1.2) é utilizada para calcular a similaridade entre dois histogramas locais de arestas.

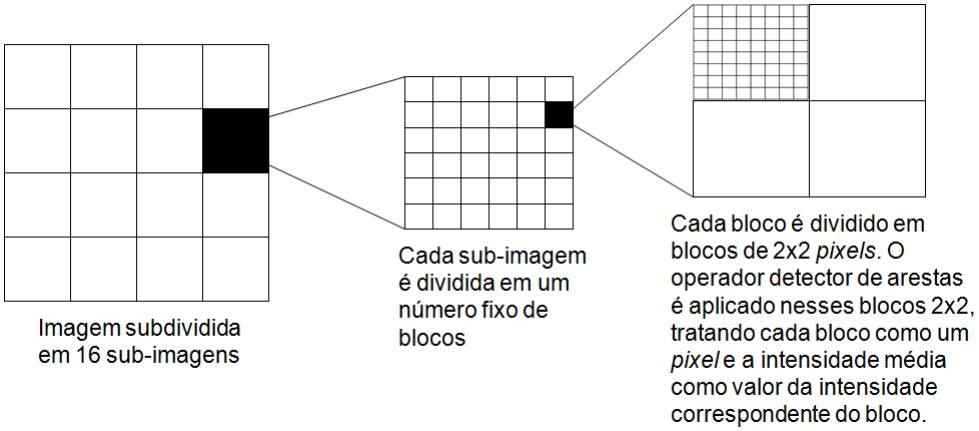

Figura 2.4: Cálculo dos histogramas locais de arestas. Ilustração baseada no artigo [Manjunath et al. 2001].

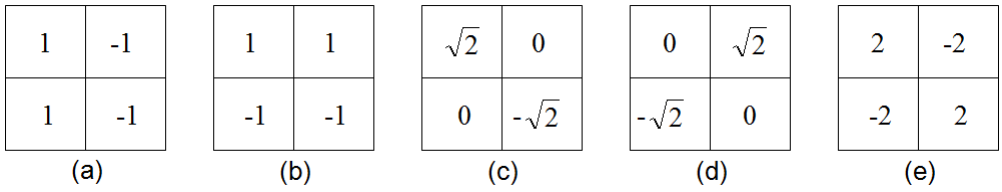

Figura 2.5: Filtros detectores de arestas: (a) vertical, (b) horizontal, (c) 45^o diagonal, (d) 135^o diagonal e (e) sem nenhuma orientação específica. Ilustração baseada no artigo [Manjunath et al. 2001].

- **CEDD - *Color and Edge Directivity Descriptor*:** Este descritor é composto [Chatzichristofis et al. 2010], pois além de características de cor, extrai

também características de textura (descrito no próximo tópico desta seção 2.1.1) da imagem. Neste descritor, pode-se considerar a imagem inteira ou um bloco dela. Para textura, é utilizado o bloco no espaço de cor YIQ. As características de textura são representadas por um histograma de 6 posições, onde 5 posições correspondem aos 5 tipos de arestas encontradas na imagem via descritor EHD e mais uma que corresponde a nenhuma aresta encontrada dos tipos indicados. Assim, dado um limiar e uma aresta, ela pode pertencer a alguma das 6 posições do histograma representando a textura da imagem.

Para cor, foi proposto um sistema *fuzzy* [Chatzichristofis e Boutalis 2007] para montar um histograma *fuzzy*. Cada bloco é processado no espaço de cor HSV. Inicialmente, é formado um histograma de 10 posições. A partir de 20 regras de inferência *fuzzy* cada posição do histograma mapeia uma cor pré-configurada: (0) *Black*, (1) *Gray*, (2) *White*, (3) *Red*, (4) *Orange*, (5) *Yellow*, (6) *Green*, (7) *Cyan*, (8) *Blue* e (9) *Magenta*. Posteriormente, o histograma é expandido em um histograma de cor de 24 posições utilizando a técnica *Coordinate Logic Filters* (CLF) para detectar arestas verticais nos três canais do espaço de cor HSV. O canal Matiz (*Hue*) é dividido em 8 áreas: *Red to Orange*, *Orange*, *Yellow*, *Green*, *Cyan*, *Blue*, *Magenta* e *Blue to Red*; o canal Saturação (*Saturation*) é dividido em 2 áreas *fuzzy* determinando a sombra da cor; o canal Brilho (*Value*) é dividido em três áreas sendo que uma delas define quando o *pixel* será preto e as outras duas, em combinação com a Saturação, quando será cinza. Baseada nesta divisão, um conjunto de 4 regras de inferência *fuzzy* são aplicadas transformando as 10 cores iniciais em um histograma de 24 posições mapeando as seguintes cores: (0) *White*, (1) *Gray*, (2) *Black*, (3) *Light Red*, (4) *Red*, (5) *Dark Red*, (6) *Light Orange*, (7) *Orange*, (8) *Dark Orange*, (9) *Light Yellow*, (10) *Yellow*, (11) *Dark Yellow*, (12) *Light Green*, (13) *Green*, (14) *Dark Green*, (15) *Light Cyan*, (16) *Cyan*, (17) *Dark Cyan*, (18) *Light Blue*, (19) *Blue*, (20) *Dark Blue*, (21) *Light Magenta*, (22) *Magenta* e (23) *Dark Magenta*.

Por fim, o histograma gerado pelo CEDD possui 144 posições pois a informação de cor é processada para todos os tipos de arestas definidas nas características de textura (6x24). A similaridade entre as duas imagens é medida utilizando o Coeficiente de Tanimoto (apresentado na subseção 2.1.2).

- **FCTH - *Fuzzy Color and Texture Histogram*:** Este descritor também é composto [Chatzichristofis et al. 2010] e se assemelha ao CEDD. O que difere é que a informação de textura no FCTH é extraída via transformada de Haar [Stollnitz et al. 1996] produzindo um histograma de 8 posições. O procedimento restante para extrair as características de cor é semelhante ao descritor CEDD e leva em consideração cada coeficiente calculado no processo de extração de características de textura. Portanto, o descritor FCTH gera um histograma de 194 posições (8x24).

O Coeficiente de Tanimoto (apresentado na subseção 2.1.2) também é utilizado como medida de similaridade no FCTH.

- **BIC - *Border-Interior Pixel Classification*:** O BIC é um descritor global de imagem que tem apresentado resultados significativos com relação a eficiência e eficácia [Stehling et al. 2002]. Este descritor é adequado para bases de imagens diversificadas e de grande volume. O primeiro passo para extração é quantizar a imagem no espaço de cor RGB em 4x4x4=64 cores. Após o passo da quantização, os *pixels* da imagem são classificados como *pixels* de borda ou *pixels* de interior. Um *pixel* é classificado como borda se pelo menos um dos quatro *pixels* vizinhos (superior, inferior, esquerda e direita) tem uma cor quantizada diferente. Já um *pixel* é classificado como interior se todos os quatro *pixels* vizinhos possuírem a mesma cor quantizada. Após a classificação dos *pixels*, são criados dois histogramas de cores: um considerando os *pixels* de borda e outro os *pixels* de interior. Assim, essa técnica da uma ideia de como os *pixels* estão distribuídos sobre a imagem fornecendo uma noção da textura da image inteira [Kimura et al. 2011]. Neste descritor é utilizada a distância dLog (apresentada na subseção 2.1.2) como medida de similaridade.

- **LCPC - *Local Color Pixel Classification*:** Ao comparar o BIC (descrito no tópico anterior) com outros descritores de cor, comumente descritos na literatura, ele apresentou melhores resultados [Penatti e da Silva Torres 2008] em termos de eficiência e eficácia. O uso do BIC permite a rápida extração de características e representação compacta das características visuais sem necessitar de grandes quantidades de espaço de armazenamento. Porém, por ser um descritor global, o vetor de características não representa a concentração de *pixels* em algumas regiões da imagem. Com isso, Kimura et al. propõe definir um descritor baseado em partições da imagem que explora a localidade de informações da distribuição dos *pixels*.

  Portanto, o LCPC [Kimura et al. 2011] é um descritor de cor, baseado em partições da imagem, que gera histogramas de cores para cada partição. Inicialmente, a imagem é uniformemente quantizada no espaço de cor RGB reduzindo em 64 o número de cores distintas. Este processo é amplamente adotado na prática e parece ser uma boa forma de reduzir o tamanho do vetor de características [Stehling et al. 2002].

  Após a quantização, a imagem é particionada em cinco partições não-sobrepostas, conforme mostra Figura 2.6. A partição central corresponde a 50% do tamanho total da imagem com o objetivo de capturar a informação presente no centro da imagem. A ideia da proposta é que, geralmente, o objeto de interesse de uma imagem fica localizado no centro dela, representando o *foreground*, e os demais objetos localizam-se no *background*. O fundo da imagem (*background*) é divido em quatro partes iguais. A Figura 2.7 mostra o tamanho das partições dada uma imagem de tamanho 100x100

pixels. As quatro partições do *background* são representadas pelas regiões superior esquerda, superior direita, inferior esquerda e inferior esquerda.

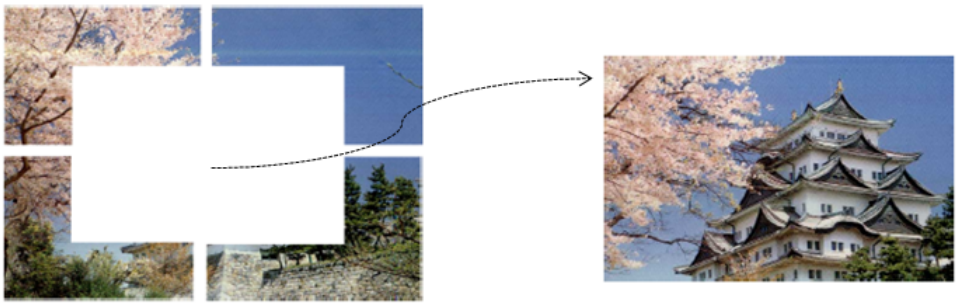

Figura 2.6: Esboço das cinco partições não-sobrepostas utilizadas pelo descritor LCPC. Ilustração baseada no artigo [Kimura et al. 2011].

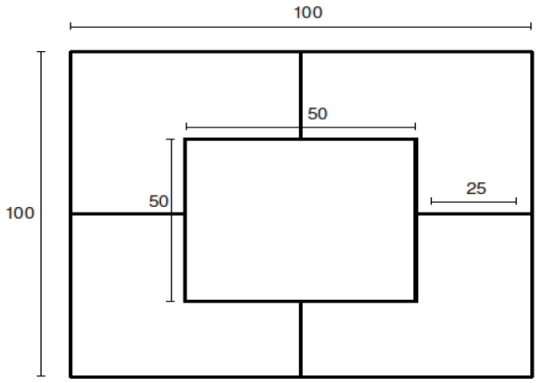

Figura 2.7: Tamanho das partições da imagem utilizadas pelo descritor LCPC. Ilustração retirada do artigo [Kimura et al. 2011].

Uma vez particionada a imagem, o processo de extração de características é igual ao utilizado pelo descritor BIC. Os *pixels* pertencentes a cada partição são classificados em borda ou interior de acordo com suas vizinhanças. Um *pixel* é classificado como borda se pelo menos um dos quatro vizinhos tem uma cor quantizada diferente ou pertence a borda da partição. Se seus quatro vizinhos possuem a mesma cor quantizada, o mesmo é classificado como interior. Assim, são criados histogramas de cores para cada partição considerando a classificação dos *pixels*. Neste caso, a imagem é representada em termos de *pixels* de borda e interior para cada cor quantizada em cada partição. Cada histograma possui 128 posições (64 para *pixels* de borda e 64 para *pixels* de interior) e consequentemente o vetor de características do LCPC possui tamanho de 640 (5x128).

A métrica de similaridade utilizada neste descritor é a distância dLog (apresentada na subseção 2.1.2) sendo que o cálculo da similaridade entre as imagens é realizado pela combinação linear de cada partição, o que garante que a partição central terá maior impacto no resultado final da similaridade.

- **LCPC + Edge - *Local Color Pixel Classification*:** Descritores baseados em cor tendem a recuperar imagens não-relevantes quando as cores do *foreground* e do *background* são similares, independente de outras características [Kimura et al. 2011]. Com o objetivo de recuperar outras imagens relevantes com cores diferentes e, ao mesmo tempo, reduzindo o número de imagens não-relevantes, Kimura et al. propuseram a combinação de características de cor do descritor LCPC com características de forma (descrito no próximo tópico desta seção 2.1.1) para enriquecer a capacidade de busca do descritor LCPC. Para isso, foi adaptado o algoritmo de detecção de bordas de Canny [Canny 1986].

  A ideia é detectar as bordas da imagem e suas direções. Dada a grande quantidade de possíveis direções de bordas, são utilizadas apenas quatro direções: 0, 45, 90 e 135 graus. Assim um histograma de 4 posições representa a quantidade de bordas em cada direção. A função de distância utilizada para avaliar a similaridade é a *Chi-square*. Ao contrário da ideia da função dLog que reduz grandes diferenças entre histogramas distintos, neste caso, qualquer grande diferença entre as bordas é importante e deve ser mantida. A similaridade entre dois histogramas $\bar{h}$ e $\bar{g}$ é dada pela Equação 2.2 [Kimura et al. 2011]:

$$D(\bar{h}, \bar{g}) = \sum_{i=0}^{i<N} \frac{(\bar{h}_i - \bar{g}_i)^2}{\bar{h}_i + \bar{g}_i} \tag{2.2}$$

  em que N representa o tamanho do histograma.

  Para considerar ambas as características de cor e forma, as mesmas são combinadas utilizando uma média ponderada das distâncias dos histogramas de cor e forma previamente calculados. O peso t é aplicado para normalizar (valores entre 0 e 1) da distância *Chi-square DQ*. O peso LCPC é igual a (1-t) e é aplicado para normalizar (valores entre 0 e 1) a distância dLog *DL*. Assim temos a distância final calculada pela Equação 2.3 [Kimura et al. 2011]:

$$D = tDQ + (1 - t)DL \tag{2.3}$$

Textura

É difícil dar uma definição precisa para a textura, no entanto, o que podemos dizer é que ela fornece informações importantes para a representação das imagens. Podemos caracterizá-la por variações locais em valores de *pixels* que se repetem de maneira regular ou aleatória ao longo do objeto ou imagem. A partir da textura conseguimos descrever o conteúdo de muitas imagens do mundo real como, por exemplo, casca de frutas, nuvens, árvores, tijolos, entre outros. Por isso, a textura é uma característica importante na

definição de semântica de alto nível para fins de recuperação de imagens por conteúdo [Liu et al. 2007].

O descritor de textura nos fornece medidas de propriedades como suavidade, rugosidade, e regularidade. Podemos citar algumas técnicas para extração de características de textura como os filtros de Gabor [Santini e Jain 1996] e as transformadas de *wavelets* [Daubechies 1990] [Santini e Gupta 2001] [Arivazhagan e Ganesan 2003]. A Figura 2.8 apresenta algumas imagens com texturas diferentes extraídas da base de texturas de [Brodatz 2012].

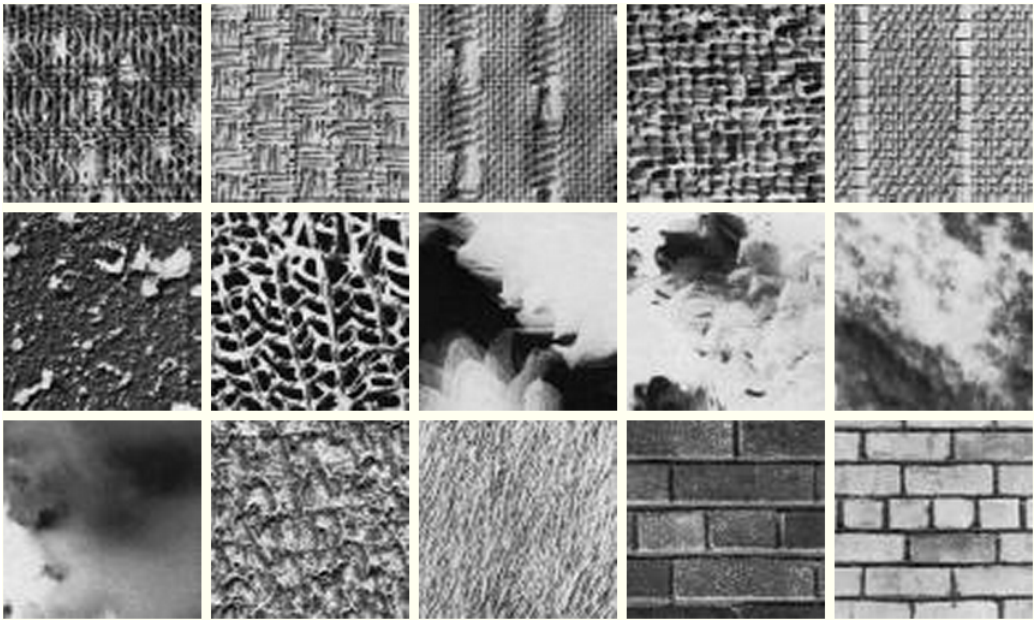

Figura 2.8: Exemplo de imagens com texturas diferentes da base de texturas de [Brodatz 2012].

Forma

As características de forma da imagem representam informações de contornos nas regiões de interesse com o objetivo de se reconhecer objetos na imagem. A principal desvantagem da extração de características de forma é que para se obter os contornos de interesse dos objetos, a imagem deve ser segmentada e o processo de segmentação depende muito do domínio da imagem levando à utilização de diferentes algoritmos de segmentação.

Os principais descritores de forma podem ser classificados em descritores baseados em região e descritores baseados em contorno. Os descritores de forma baseados em região expressam a distribuição dos *pixels* como uma região de um objeto podendo descrever objetos complexos consistindo de múltiplas regiões desconexas e/ou objetos simples com ou sem buracos, considerando as formas internas do objeto. A Figura 2.9 mostra exemplos de imagens nas quais os descritores de forma baseado em regiões poderiam ser aplicados

[Bober 2001]. Como exemplo desses tipos de descritores podemos citar os Descritores de Fourier [Sonka et al. 1998], *Curvature Scale Space* [Abbasi et al. 2000] entre outros.

Figura 2.9: Exemplo de formas para as quais um descritor de forma baseado em região é aplicável. Ilustração retirada do artigo [Bober 2001].

Já os descritores baseados em contornos expressam as propriedades da forma pelo seu esboço (contorno) considerando as formas mais externas do objeto. Objetos para os quais características de forma estão contidas no contorno são descritas eficientemente por este tipo de descritor, conforme mostrado na Figura 2.10 [Bober 2001]. Como exemplo desses tipos de descritores podemos citar Coeficiente de Compacidade [Sonka et al. 1998], Momentos de Zernike [Kotoulas e Andreadis 2005], entre outros.

Figura 2.10: Exemplo de formas para as quais um descritor de forma baseado em contorno é aplicável. Ilustração retirada do artigo [Bober 2001].

Podemos exemplificar a utilização de descritores de forma em uma aplicação de apoio ao diagnóstico de câncer de mama. A Figura 2.11 mostra um exemplo da extração de contorno de um tumor de mama em uma mamografia. A partir deste contorno podemos extrair informações como a área, perímetro, coeficiente de compacidade entre outros.

A área (A) é obtida contando o número de *pixels* que formam a região do contorno. O perímetro (P) é calculado considerando o número de *pixels* conexos que constituem o contorno da região. Por fim, o coeficiente de compacidade (cc) [Sonka et al. 1998] que descreve o quanto a forma se aproxima de uma circunferência com a mesma área é obtida pela Equação 2.4:

$$cc = \frac{P^2}{4\pi A} \tag{2.4}$$

2.1.2 Medidas de Similaridade

No processo de recuperação de imagens por conteúdo são utilizadas as medidas de similaridade para encontrar as imagens da base mais similares a uma determinada imagem de referência fornecida pelo usuário. Essas medidas são funções de distância que aplicadas aos vetores de características de cada imagem retornam um valor real positivo utilizado

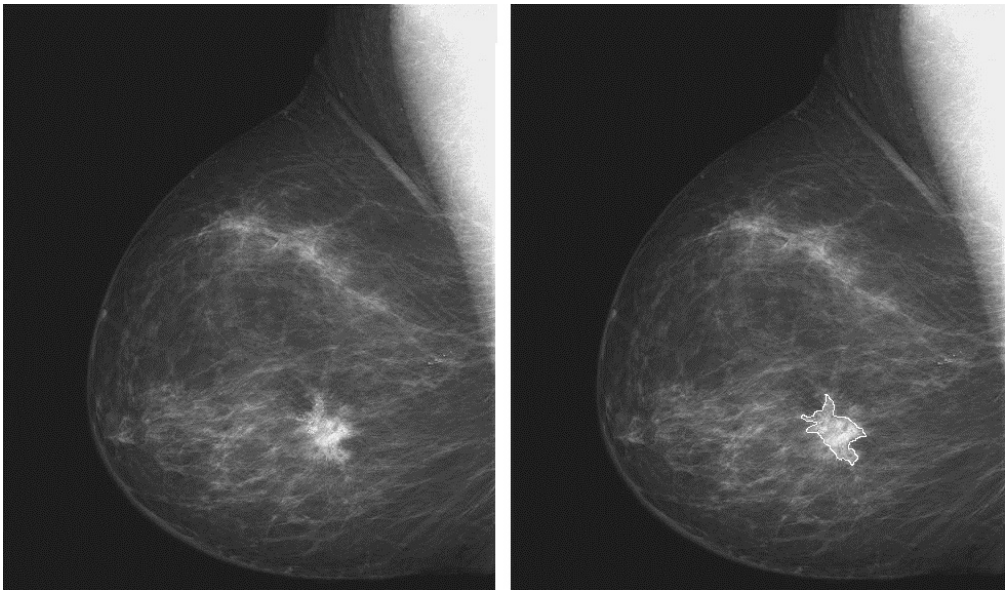

Figura 2.11: Exemplo da extração do contorno de um tumor em uma mamografia digital.

para ordenar a lista de imagens retornadas. Quanto maior o valor calculado menos similar a imagem analisada é da imagem de referência.

A escolha da melhor função de distância a ser utilizada para o cálculo de similaridade dos vetores deve levar em consideração os descritores utilizados na geração desses vetores.

Modelos matemáticos utilizando características comuns de várias distâncias diferentes, são conhecidos como espaços métricos. Assim, para descrever uma função de distância é utilizado um modelo matemático chamado espaço métrico [Jacobs et al. 2000] [Kutz et al. 2003]. O espaço métrico é representado pelo par $<W,\ d>$, em que W é um conjunto de elementos (vetores de características) e d é uma função de distância, denominada métrica, que satisfaz os seguintes axiomas para qualquer elemento $\bar{x},\ \bar{y},\ \bar{z} \in W$:

- Identidade: $d(\bar{x},\bar{y}) \geq 0$ e $d(\bar{x},\bar{y}) = 0$, sse $\bar{x} = \bar{y}$
- Simetria: $d(\bar{x},\bar{y}) = d(\bar{y},\bar{x})$
- Não-negatividade: $0 \leq d(\bar{x},\bar{y}) < \infty$
- Desigualdade triangular: $d(\bar{x},\bar{z}) \leq d(\bar{x},\bar{y}) + d(\bar{y},\bar{z})$

As métricas mais conhecidas e utilizadas são aquelas da família Minkowski ou métricas L_p. Essas métricas são definidas pela Equação 2.5 variando o parâmetro $p \in \mathbb{R} \mid p \geq 1$, sendo que $\bar{x}$ e $\bar{y}$ são vetores e n é o número de dimensões desse vetores. Para p=1 temos a distância Manhattan ou *City Block* (L_1), p=2 a distância Euclidiana (L_2) e p=∞ a distância *Infinity* ou Chebychev (L_∞).

$$d(\bar{x},\bar{y}) = \sqrt[p]{\sum_{i=1}^{n} |\bar{x}_i - \bar{y}_i|^p} \tag{2.5}$$

A Figura 2.12 mostra a abrangência geométrica dessas funções em um espaço bidimensional. A distância L_1 é como se o deslocamento de $\bar{x}$ até $\bar{y}$ fosse possível por meio de segmentos lineares paralelos ou perpendiculares entre si, da forma como ocorre nas ruas de uma cidade (por isso, o nome *City Block*). Já a distância L_2 define o lugar geométrico de todos os pontos equidistantes do ponto que se deseja calcular a distância, ou seja, uma circunferência no espaço 2D com centro no ponto que se deseja calcular a distância. Por fim, a distância L_∞ entre $\bar{x}$ e $\bar{y}$ é determinada unicamente pelo valor representado no eixo no qual a diferença entre as coordenadas do ponto é maior.

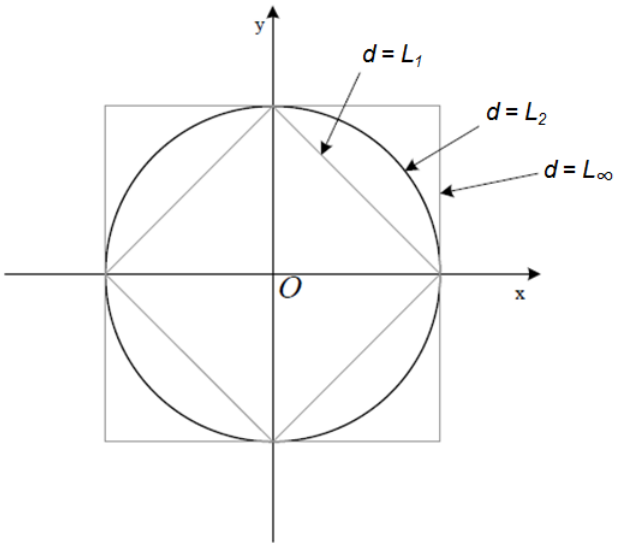

Figura 2.12: Abrangência geométrica das funções de distâncias no espaço bidimensional, em que d representa a distância (L_1, L_2 ou L_∞) até a origem O. Ilustração retirada do artigo [Rohani e Nugroho 2008].

Outro modelo matemático também muito conhecido é o espaço vetorial. Nesse modelo definimos a distância cosseno entre vetores que quanto menor for o ângulo entre os vetores de características $\bar{x}$ e $\bar{y}$, maior será o grau de similaridade de $\bar{x}$ e $\bar{y}$. Além disso, essa medida não satisfaz a propriedade da desigualdade triangular e $d(\bar{x},\bar{y}) = 1$. A Equação 2.6 apresenta o cálculo da distância cosseno entre os vetores $\bar{x}$ e $\bar{y}$ cuja a dimensão seja n:

$$d(\bar{x}, \bar{y}) = \frac{\sum_{i=1}^{n} \bar{x}_i \bar{y}_i}{\sqrt{\sum_{i=1}^{n} \bar{x}_i^2 \sum_{i=1}^{n} \bar{y}_i^2}} \tag{2.6}$$

O cosseno é muito utilizado para medir a distância entre vetores cujos elementos representam a frequência de ocorrência com que um termo de um dicionário aparece em um documento (abordagem *bag-of-words*, descrita na seção 2.3) ou com que uma palavra visual de um dicionário aparece em uma imagem (abordagem *bag-of-visual-features*, descrita na seção 2.3). Nesta última representação, duas imagens com conteúdo muito semelhante podem ter uma diferença significativa entre os vetores simplesmente porque um é muito maior do que o outro. Assim, a distribuição relativa das palavras visuais pode ser idêntica nas duas imagens, mas a freqüência absoluta de um vetor pode ser muito maior do que

o de outro. A distância cosseno evita este problema pois independe da magnitude dos vetores, uma vez que o cálculo da distância considera o cosseno do ângulo θ entre os dois vetores, conforme apresentado na Figura 2.13 [Manning et al. 2008].

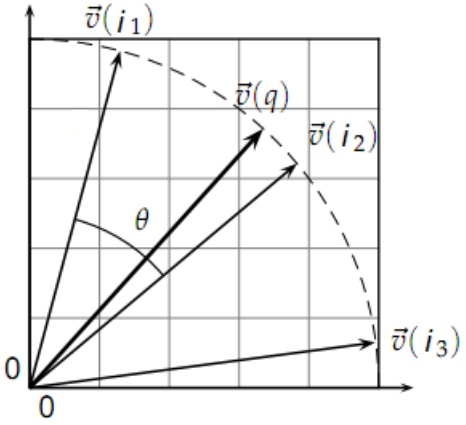

Figura 2.13: Ilustração do cálculo da distância cosseno entre duas imagens, $d(\vec{v}(i_1),\vec{v}(i_2))$=cos$\theta$. Ilustração retirada do artigo [Manning et al. 2008].

Outra medida de similaridade encontrada na literatura utiliza o Coeficiente não-binário de Tanimoto para calcular a distância entre vetores de características [Chatzichristofis et al. 2010]. A distância entre duas imagens cujo os vetores são $\bar{x}_i$ e $\bar{x}_j$ respectivamente é representada pela Equação 2.7:

$$d(\bar{x}_i, \bar{x}_j) = \frac{\bar{x}_i^T \bar{x}_j}{\bar{x}_i^T \bar{x}_i + \bar{x}_j^T \bar{x}_j - \bar{x}_i^T \bar{x}_j} \tag{2.7}$$

em que $\bar{x}^T$ é o vetor transposta do vetor $\bar{x}$.

Na congruência absoluta dos vetores, a distância utilizando o Coeficiente de Tanimoto tem valor igual a 1, enquanto que quando houver o desvio máximo dos vetores, o valor tenderá a 0 [Chatzichristofis et al. 2010].

A função de distância dLog calcula a diferença entre o logaritmo dos elementos presente nos vetores de características com o objetivo de reduzir o efeito negativo quando alguns elementos do vetor possui alguns valores muito altos e outros valores muito baixos. Estes valores elevados podem dominar a diferença entre dois vetores. Assim, a função dLog é utilizada para minimizar essa distorção [Kimura et al. 2011]. Ela é definida pelas Equações 2.8 e 2.9:

$$d(\bar{a}, \bar{b}) = \sum_{i=0}^{i<n} |f(\bar{a}_i) - f(\bar{b}_i)| \tag{2.8}$$

$$f(x) = \begin{cases} 0, & \text{se } x = 0 \\ 1, & \text{se } 0 < x \leq 1 \\ \lceil \log_2 x + 1 \rceil, & \text{caso contrário} \end{cases} \tag{2.9}$$

Na Equação 2.8, $\bar{a}$ e $\bar{b}$ são vetores de características de tamanho n. O valor $\bar{a}_i$ representa o valor da posição i do vetor $\bar{a}$ e o valor $\bar{b}_i$ representa o valor da posição i do vetor $\bar{b}$.

2.1.3 Operadores de Similaridade

As consultas realizadas em banco de dados tradicionais que manipulam dados numéricos e textuais são exatas, baseadas em igualdade ou em ordem total. Já as consultas para recuperação de imagens por conteúdo são baseadas em similaridade de características. Dessa forma, após extrair as características de todas as imagens da base de consulta e definir uma medida de similaridade (apresentadas na seção anterior) baseando-se nos vetores de características dessas imagens, é necessário definir o operador de consulta a ser utilizado. Os operadores de consulta geralmente são definidos como um elemento de busca (imagem de referência) e uma restrição baseada na proximidade (distância) em relação aos elementos de um conjunto (imagens da base de consulta). A seguir são apresentados os dois operadores de consulta por similaridade mais conhecidos: *Range Query* (RQ) e *K-Nearest Neighbor Query* (KNNQ) [Zezula et al. 2006].

Range Query (RQ)

Este operador recupera todos os elementos que possuem uma distância menor ou igual a uma distância r pré-definida, com relação a um elemento de referência q conforme mostra a Figura 2.14. Seja $\mathbb{X}$ um domínio de dados, o operador RQ recupera todo elemento x de um conjunto $\mathrm{X} \subseteq \mathbb{X}$ que se encontra a até uma distância máxima r (raio de busca) do elemento q de referência, onde $q \in \mathbb{X}$. Formalmente temos a Equação 2.10:

$$RQ(q, r) = \{x \in X | d(q, x) \leq r\} \tag{2.10}$$

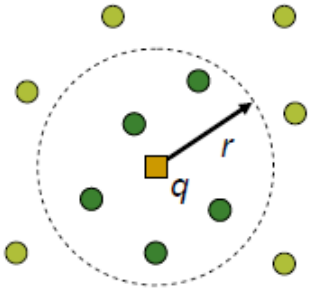

Figura 2.14: Consulta por abrangência com centro de referência q e raio de busca r. Ilustração baseada no livro [Zezula et al. 2006].

***K-Nearest Neighbor Query* (KNNQ)**

O operador KNNQ recupera k elementos mais próximos ao elemento de referência q como ilustrado na Figura 2.15. Diferentemente do operador RQ não há a necessidade de determinar um raio de busca r, o que pode ser uma tarefa difícil sem um prévio conhecimento da distribuição do conjunto de dados e da função de distância.

Assim, uma consulta aos k-vizinhos mais próximos recupera os k elementos do conjunto de dados $X \subseteq \mathbb{X}$ mais próximos ao elemento de consulta $q \in \mathbb{X}$. Formalmente temos a Equação 2.11:

$$KNNQ(q,k) = \{R \subseteq X, |R| = k \wedge \forall x \in R, y \in X - R : d(q,x) \leq d(q,y)\} \quad (2.11)$$

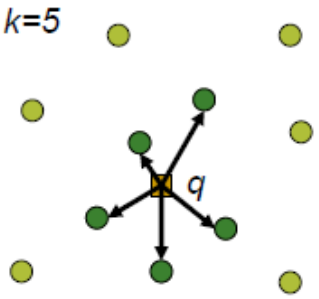

Figura 2.15: Consulta dos k-vizinhos mais próximos a partir do centro de consulta q e k=5. Ilustração baseada no livro [Zezula et al. 2006].

2.2 Avaliação dos sistemas de recuperação

Na área de Recuperação de Informação ao realizar uma consulta de um determinado documento pode-se obter como resultado vários documentos similares. Assim, necessita-se de técnicas para medir a precisão da relevância desses documentos similares de acordo com a necessidade do usuário. Segundo Baeza [Baeza-Yates e Ribeiro-Neto 1999], este tipo de avaliação é conhecido por Avaliação de Desempenho de Recuperação. As medidas mais utilizadas para a avaliação de sistemas de recuperação de informação são a **Precisão** e **Revocação** [Manning et al. 2008]. Essas medidas também são muito utilizadas para medir o desempenho de sistemas de recuperação de imagem por conteúdo que ao invés considerarem documentos, processam imagens como objetos de consulta.

A **Precisão** corresponde à capacidade do sistema recuperar somente os documentos relevantes. Assim, calcula-se a **Precisão** (P) pela razão dos documentos recuperados e que são relevantes ($|R_{rec,rel}|$) e o total de documentos recuperados ($|D|$) conforme mostra

a Equação 2.12:

$$P = \frac{|R_{rec,rel}|}{|D|} \tag{2.12}$$

Já a **Revocação** corresponde à capacidade do sistema recuperar todos os documentos relevantes. O cálculo da **Revocação** (R) é a razão entre os documentos recuperados e que são relevantes ($|R_{rec,rel}|$) e o total de documentos relevantes ($|R_{rel}|$) conforme mostra a Equação 2.13:

$$R = \frac{|R_{rec,rel}|}{|R_{rel}|} \tag{2.13}$$

Para facilitar a análise de desempenho do Sistema de Recuperação, são geradas as curvas de **Precisão** x **Revocação** que representa a precisão para vários níveis de revocação. Com essas curvas pode-se comparar o desempenho entre diversos sistemas de recuperação. A Figura 2.16 mostra as curvas de precisão x revocação de dois sistemas de recuperação, S1 e S2. Neste exemplo, o sistema S1 é mais preciso que S2 em baixas revocações porém a partir de um certo nível de revocação (próximo de 30%), S2 passa a ser mais preciso que S1.

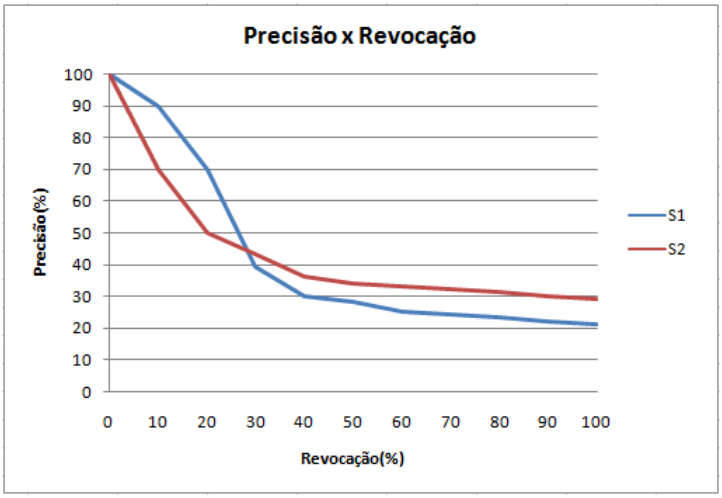

Figura 2.16: Exemplo da curva de precisão x revocação de dois sistemas de recuperação (S1 e S2).

Conforme verificado na Figura 2.16 não é fácil determinar qual sistema de recuperação é mais preciso na média. Assim, é necessário a utilização de um valor médio único para representar a qualidade de recuperação de um determinado sistema. Nas próximas subseções é mostrada uma maneira de calcular esse valor médio único.

2.2.1 Precisão Média (*Average Precision*)

A precisão média (*AveP - Average Precision*) calcula a média dos valores de precisões (*P(k)*) obtidos para um determinado sistema de recuperação, em que k representa o

número de documentos recuperados. Essa medida representa o desempenho do sistema na recuperação dos documentos relevantes. A Equação 2.14 mostra como se calcula o ($AveP$):

$$AveP = \frac{\sum_{k=1}^{|D|} P(k)rel(k)}{|R_{rel}|} \tag{2.14}$$

em que $|D|$ representa o total de documentos recuperados, $|R_{rel}|$ representa o total de documentos relevantes e *rel(k)* indica a relevância do documento k recuperado, sendo 1 para relevante e 0 para não relevante.

2.2.2 Média dos Valores de Precisão Média (*Mean Average Precision*)

A média dos valores de precisão média (*MAP - Mean Average Precision*) mede o desempenho médio de um sistema de recuperação para Q consultas realizadas. A Equação 2.15 apresenta o cálculo dessa medida:

$$MAP = \frac{\sum_{q=1}^{Q} AveP_q}{Q} \tag{2.15}$$

em que $AveP_q$ é a precisão média para a consulta q e Q é o número de consultas realizadas. Entre as medidas de avaliação, o MAP tem demonstrado ter boa discriminação e estabilidade [Manning et al. 2008].

2.2.3 Ganho Acumulativo Descontado Normalizado (*Normalized Discount Cumulative Gain*)

A medida de avaliação de Ganho Acumulativo Descontado (*DCG - Discount Cumulative Gain*) vem sendo muito utilizada para avaliar o desempenho de algoritmos de busca na Web e sistemas de recuperação de informação. Nessa medida, é levada em consideração a posição do item recuperado na lista de *ranking*, pois resultados corretos que são recuperados no início da lista tem mais valor do que os que são recuperados no final da mesma. Isso se deve ao fato de que é menos provável que o usuário avalie instâncias mais distantes dos primeiros resultados obtidos em uma busca [Järvelin e Kekäläinen 2000] [Shilane et al. 2004].

Com base na lista de *ranking*, considerando uma consulta, o DCG acumula o ganho de cada resposta baseando-se na posição dessa no *ranking* descontando do valor acumulado os itens relevantes que aparecem em posições mais baixa no *ranking*. A Equação 2.16

mostra como é obtido o valor DCG_k para até uma posição k do *ranking*:

$$DCG_k = rel_1 + \sum_{i=2}^{k} \frac{rel_i}{\log_2 i} \tag{2.16}$$

em que $rel_i \in \{0, 1\}$, considerando aplicações cuja a relevância dos objetos é binária, 0 corresponde a uma resposta irrelevante e 1 a uma resposta relevante.

Assim, para tornar possível a comparação do desempenho de diferentes algoritmos, os valores do *DCG*, obtidos para cada algoritmo, devem ser normalizados [Manning et al. 2008] [Shilane et al. 2004]. Portanto, a Equação 2.17 mostra a normalização da medida *DCG*, a qual chamamos de Ganho Acumulativo Descontado Normalizado (*NDCG - Normalized Discount Cumulative Gain*):

$$NDCG_k = \frac{DCG_k}{IDCG_k}, IDCG_k = 1 + \sum_{i=2}^{k} \frac{1}{\log_2 i} \tag{2.17}$$

em que o *IDCG* corresponde ao valor ideal do *DCG*, obtido com um *ranking* perfeito.

Os valores de $NDCG_k$ variam no intervalo de [0, 1]. Os resultados do *NDCG* de todas as consultas podem ser totalizados em uma média aritmética para que esse resultado seja utilizado como a média de desempenho do algoritmo utilizado.

2.3 Bag-Of-Visual-Features

O *bag-of-visual-features*[1] é um modelo de representação das características visuais de um conjunto de imagens com intuito de otimizar a recuperação e ou classificação nesse conjunto reduzindo a diferença semântica entre as características de baixo nível e o conteúdo visual da imagem. É uma técnica que tem obtido bons resultados quando utilizado para reconhecimento de objetos, principalmente devido à sua robustez a vários tipos de variações e oclusão. Um dos trabalhos pioneiros nessa abordagem foi proposto por Sivic e Zisserman [Sivic e Zisserman 2003], que apresenta a técnica como uma abordagem para recuperar todas as ocorrências de um objeto em cenas (*frames*) de um determinado vídeo. Para isso, os objetos são representados como um conjunto de descritores invariantes a escala, rotação, translação, iluminação e oclusão parcial.

Desde então outros trabalhos foram propostos para os mais diversos domínios utilizando esta abordagem. Csurka et al. [Csurka et al. 2004] utilizam *bag-of-visual-features* com o objetivo de encontrar um processo que seja genérico para lidar com diversos tipos de objetos e ao mesmo tempo tratar as variações de iluminação, visualização, rotação e oclusão, típicos de cenas do mundo real. Batista et al. [Batista et al. 2009] utilizam esta

[1] Alguns trabalhos utilizam a denominação *bag-of-keypoints*, *bag-of-features*, *bag-of-visual-words* ou *bag-of-words* referindo-se ao mesmo método.

abordagem para a detecção de edificações em fotografias históricas e Lopes et al. [Lopes et al. 2009b] [Lopes et al. 2009a] para a detecção de nudez.

Esta abordagem foi inspirada na técnica chamada *bag-of-words* [Dumais et al. 1998] [Baeza-Yates e Ribeiro-Neto 1999] utilizada em recuperação de informações textuais (RI - Recuperação da Informação), na qual cada documento é representado por um vetor das palavras-chaves que ocorrem nele. Esses vetores são criados a partir de um dicionário de palavras.

No método *bag-of-visual-features*, o dicionário de palavras é chamado de dicionário de palavras visuais (*codebooks*) que corresponde ao agrupamento das características locais dos pontos de interesse das imagens. Em muitos trabalhos da literatura, cada grupo (*cluster*) de pontos de interesse possui o seu representante, o centróide, que é tratado como uma palavra visual no dicionário. As características locais, utilizadas para construir o dicionário, são extraídas das imagens utilizando métodos de detecção de pontos de interesse [Lindeberg 1993] [Lazebnik et al. 2003]. Esses pontos de interesse, por sua vez, são representados por descritores de características como *Invariant Feature Transform* (SIFT) [Lowe 2004], *Principal Component Analysis* (PCA-SIFT) [Ke e Sukthankar 2004] e o *Speeded Up Robust Features* (SURF) [Bay et al. 2006].

Por fim, são criados os histogramas de frequência das palavras visuais para cada uma das imagens da base de dados [Yang et al. 2007]. A partir destes histogramas é calculada a similaridade entre as imagens da base de dados e a imagem da consulta utilizando algum operador de similaridade.

A Figura 2.17 ilustra todo o processo para obtenção do dicionário de palavras visuais e para a descrição das imagens via histograma de frequência.

Visando melhorar a eficiência computacional e reduzir a utilização de memória no processo de reconhecimento de objetos utilizando o *bag-of-visual-features*, alguns trabalhos foram desenvolvidos. A alternativa foi compactar o dicionário de palavras visuais mantendo o seu poder discriminativo. Winn et al. [Winn et al. 2005] e Wang et al. [Wang et al. 2008] desenvolveram métodos semelhantes para compactar o dicionário de palavras visuais.

Além disso, Nister e Stewenius [Nister e Stewenius 2006] propuseram um método que gera uma árvore do dicionário de palavras visuais utilizando agrupamento hierárquico. Essa árvore define, de forma integrada, o dicionário de palavras visuais e uma estratégia de busca melhorando o mecanismo de indexação para o processo de recuperação. Jégou et al. [Jégou et al. 2010] também propuseram a criação de uma árvore hierárquica, porém, segundo o autor, o processo de criação dessa árvore difere do de Nister e Stewenius por ser mais custoso e ser *bottom-up* ao invés de *top-down*. Ambos os trabalhos obtiveram um dicionário de palavras compacto e discriminativo.

Assim, de uma forma geral, podemos listar as seguintes fases da abordagem *bag-of-visual-features*:

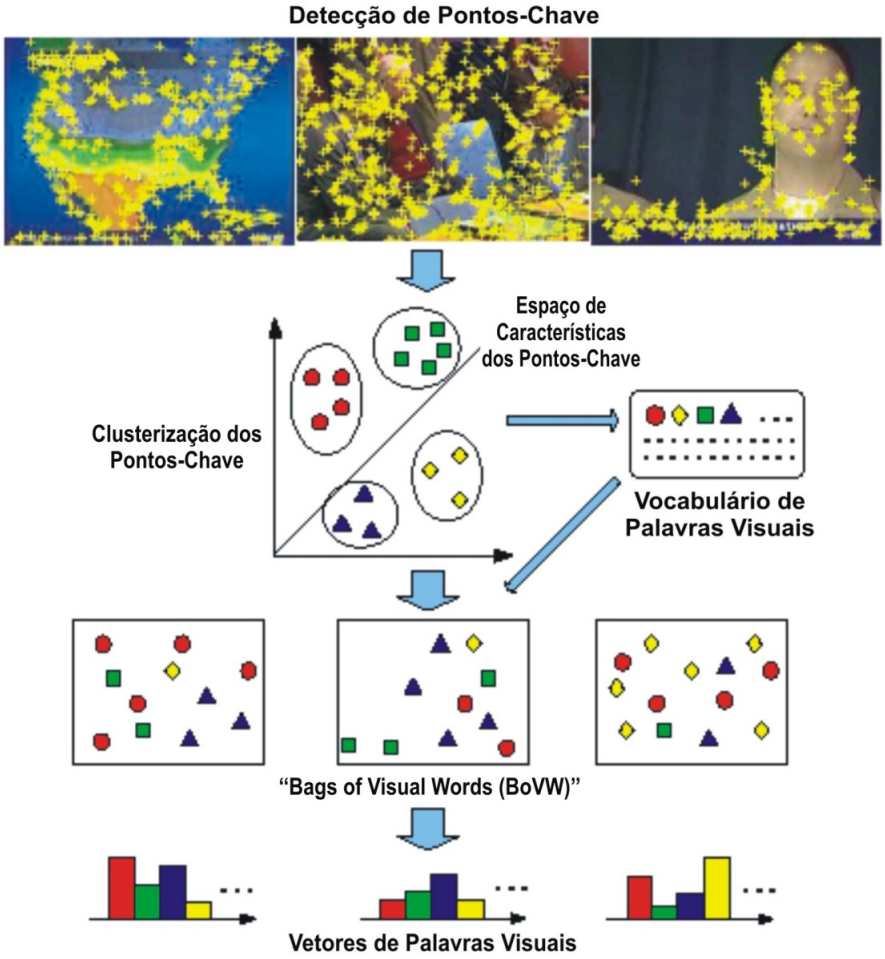

Figura 2.17: Visão geral do processo *bag-of-visual-features*. Ilustração baseada no artigo [Yang et al. 2007].

1. Extração de Características;
2. Construção do Dicionário de Palavras Visuais;
3. Construção dos Histogramas de Palavras Visuais;
4. Busca por Similaridade.

Nas próximas subseções são apresentadas e detalhadas as principais fases da abordagem *bag-of-visual-features* considerando o processo de recuperação de imagens por conteúdo.

2.3.1 Extração de Características

Esta fase é responsável por definir os vetores de características utilizados na construção do dicionário de palavras visuais. Para isso temos que detectar os pontos de interesse das imagens e extrair as características das regiões destes pontos. O ideal é que os descritores de características utilizados sejam invariáveis às transformações das imagens

(mudança de iluminação, escala, rotação, ruídos entre outros) e ricos o suficiente para extrair características discriminatórias [Csurka et al. 2004].

Existem várias técnicas de detecção de pontos de interesse ou pontos-chave de uma imagem, por exemplo: detectores de amostragem multi-escala *Harris-Laplace* [Lazebnik et al. 2003] e *Laplacian of Gaussian* [Lindeberg 1993] ou ainda detectores aleatórios [Nowak et al. 2006]. A Figura 2.18 mostra exemplo dos pontos-chave detectados em uma imagem utilizando as técnicas citadas.

Figura 2.18: Imagens destacando os pontos de interesse detectados respectivamente pelos seguintes detectores: *Harris-Laplace*, *Laplacian of Gaussian* e Aleatório. Imagens retiradas do artigo [Nowak et al. 2006].

Após a detecção dos pontos de interesse, os mesmos são descritos a partir de descritores de características visando a montagem de vetores de características. Um dos descritores mais utilizados na literatura para esse fim é o SIFT (*Scale Invariant Feature Transform*) que também realiza a detecção dos pontos de interesse.

O SIFT é uma técnica, definida por David Lowe [Lowe 2004], que permite a detecção e extração de descritores locais, razoavelmente invariáveis a mudanças de iluminação, ruído de imagem, rotação, escala e pequenas mudanças de perspectiva. Estes descritores podem ser utilizados para fazer a correspondência de diferentes visões de um objeto ou cena.

Para extrair as características de uma imagem através do SIFT, são necessárias quatro etapas. Segue uma descrição de cada etapa, maiores detalhes podem ser encontrados no trabalho de [Lowe 2004].

1. **Detecção de extremos:** Nesta etapa são detectados os extremos (máximos e mínimos) da imagem realizando a sua convolução com a função de Diferença Gaussiana (DoG - *Difference of Gaussian*) em várias escalas. Estes extremos correspondem a pontos de interesse locais que são invariantes a mudanças de escala da imagem. O espaço da escala é definido como uma função, $L(x,y,\sigma)$, obtida pelo resultado da convolução da imagem de entrada $I(x,y)$ com a função Gaussiana de escala variável $G(x,y,\sigma)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \tag{2.18}$$

em que $*$ é a operação de convolução em x e y, e

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \tag{2.19}$$

Podemos perceber que este filtro varia a escala de acordo com o parâmetro σ. A Figura 2.19 mostra exemplos de aplicação dos filtros descritos variando σ.

Figura 2.19: Exemplo de imagens após a aplicação do filtro Gaussiano variando σ.

A função DoG é dada pela diferença de imagens filtradas em escalas próximas separadas por uma constante k. A função DoG é definida pela Equação 2.20:

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \tag{2.20}$$

A Figura 2.20 mostra o resultado da função DoG aplicada as imagens da Figura 2.19.

Figura 2.20: Filtro DoG para as imagens apresentadas na Figura 2.19.

A criação das imagens *D(x,y,σ)* a partir da diferença das imagens convolucionadas é esboçado na Figura 2.21. Inicialmente a imagem é convolucionada incrementalmente aplicando o filtro Gaussiano variando σ de acordo com o fator multiplicativo k (primeira oitava, à esquerda da figura). As imagens obtidas são subtraídas conforme a Equação 2.20 representando a diferença de Gaussianas (primeira oitava, à direita da figura). O processo apresentado gera o que é chamado de uma oitava. Este processo é repetido para um número desejado de oitavas. Cada oitava representa um

conjunto de imagens $L(x,y,k\sigma)$ e $D(x,y,\sigma)$ para a imagem reescalada com diferentes amostragens. Portanto, para gerar a próxima escala, a imagem da diferença de Gaussianas que possui o dobro do valor σ inicial do nível anterior é escolhida e reduzida à metade de sua resolução utilizando a interpolação bilinear. Esta será a primeira imagem da segunda oitava, que obedecerá ao mesmo procedimento da primeira oitava.

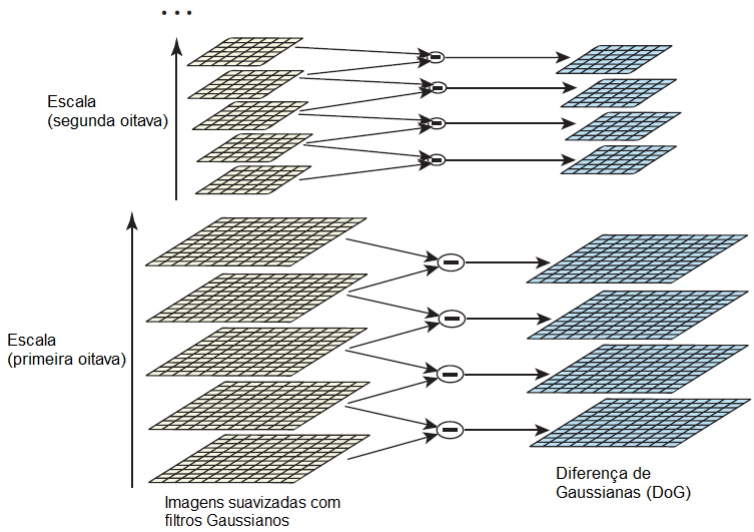

Figura 2.21: Esquema de criação das imagens $D(x,y,\sigma)$. Ilustração baseada no artigo [Lowe 2004].

A Figura 2.22 apresenta o esboço da criação das imagens $D(x,y,\sigma)$ utilizando imagens reais.

Após a criação das oitavas, é realizada a detecção de extremos. Os extremos são dados por valores locais de máximo ou mínimo para cada $D(x,y,\sigma)$. Assim, para detectar um máximo ou mínimo local de $D(x,y,\sigma)$, cada ponto é comparado com seus oito vizinhos mais próximos na imagem corrente e nove vizinhos na escala acima e abaixo (conforme mostra a Figura 2.23). O ponto é selecionado na imagem de análise se ele for maior ou menor que todos seus vizinhos, sendo considerado um potencial ponto de interesse.

Na próxima etapa é definida a localização exata dos pontos de interesse descartando os pontos de interesse instáveis.

2. **Localização dos pontos de interesse:** Esta etapa tem como objetivo definir quais pontos de interesse candidatos serão escolhidos para serem descritos na última etapa. Além disso, para os pontos escolhidos será definindo a sua localização, escala e a razão das curvaturas principais. Esta razão servirá para ajudar na rejeição dos pontos que possuem baixo contraste ou aqueles que estão localizados em bordas mal definidas.

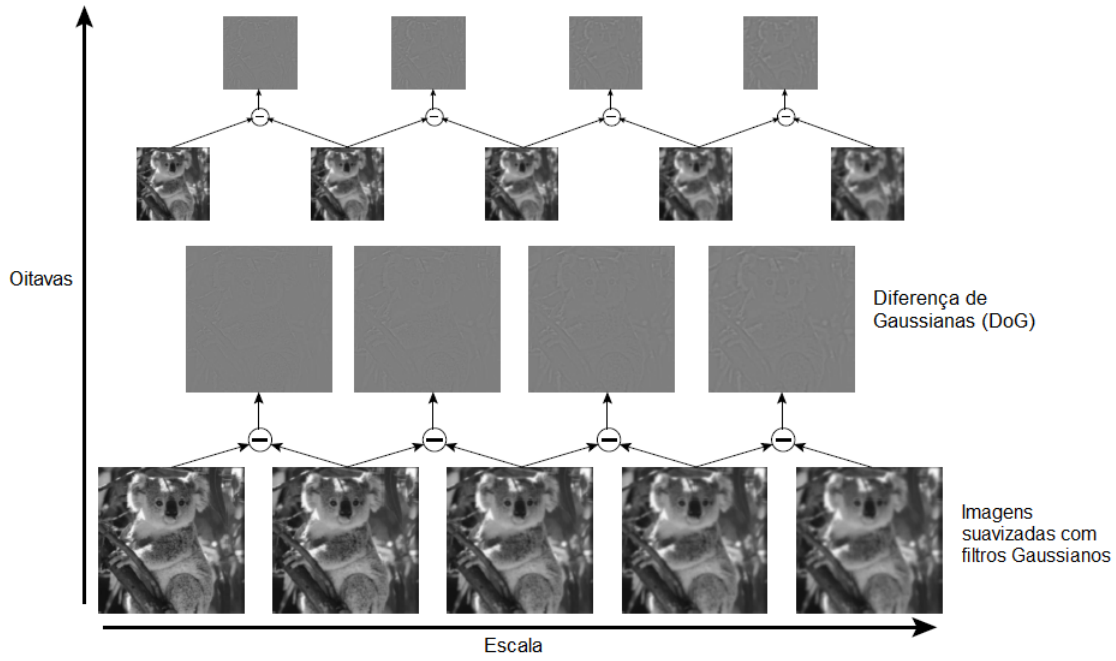

Figura 2.22: Esquema de criação das imagens *D(x,y,σ)*. Ilustração baseada no artigo [Almeida et al. 2009].

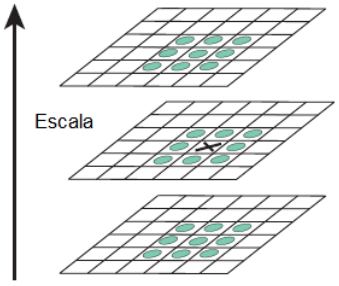

Figura 2.23: Extremos máximos e mínimos das imagens geradas pela Diferença de Gaussianas que são detectados por comparação de um *pixel* (marcado com X) com os seus 26 vizinhos em regiões de 3x3 nas escalas atuais e adjacentes (marcado com círculos). Ilustração baseada no artigo [Lowe 2004].

Uma vez que um ponto de interesse candidato foi encontrado, o próximo passo é ajustar a sua precisão de localização. Para cada ponto analisado é utilizada uma expansão de Taylor da função *D(x,y,σ)* transladada, conforme mostra a Equação 2.21, de modo que a origem desta expansão esteja localizada no ponto de origem:

$$D(\bar{x}) = D + \frac{\partial D^T}{\partial \bar{x}}\bar{x} + \frac{1}{2}\bar{x}^T\frac{\partial^2 D}{\partial \bar{x}^2}\bar{x} \tag{2.21}$$

em que

$$D = D(x, y, \sigma) \tag{2.22}$$

$$D(\bar{x}) = D(x + x', y + y', \sigma + \sigma') \tag{2.23}$$

Esta equação deve ser entendida da seguinte maneira, D e suas derivadas são avaliadas a partir do ponto analisado e $\bar{x} = (x', y', \sigma')^T$ é o *offset* em relação a este ponto. Ou seja, D é o valor da função $D(x,y,\sigma)$ no ponto avaliado, $\bar{x}$ é o *offset* em relação a este ponto e $D(\bar{x})$ é a aproximação do valor de $D(x,y,\sigma)$ interpolado para um ponto transladado com *offset* $\bar{x}$.

Assim, a localização do extremo, $\hat{x}$, é determinada fazendo a derivada segunda da Equação 2.21 com relação a $\bar{x}$ e igualando o resultado a zero, conforme mostra a Equação 2.24:

$$\hat{x} = -\frac{\partial^2 D^{-1}}{\partial \bar{x}^2}\frac{\partial D}{\partial \bar{x}} \tag{2.24}$$

O valor da função no extremo, $D(\hat{x})$, é utilizado para se rejeitar extremos instáveis com baixo contraste. Substituindo-se a Equação 2.24 na Equação 2.21 obtemos a Equação 2.25:

$$D(\hat{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial \bar{x}}\hat{x} \tag{2.25}$$

Lowe aconselha que se rejeite todos os extremos com valores de $|\mathrm{D}(\hat{x})|$ inferiores a 0.03 (assumindo-se que os pixels da imagem estejam entre [0,1]).

Um pico mal definido em DoG terá grande curvatura principal ao longo da borda, porém pequena curvatura em sua direção perpendicular. As curvaturas principais podem ser computadas a partir da matriz Hessiana 2x2, $\mathbf{H}$, computada na localização e escala do ponto:

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix} \tag{2.26}$$

Os autovalores de $\mathbf{H}$ são proporcionais as principais curvaturas de D. Porém, não será necessário computar os autovalores pois o que se busca é a razão entre as curvaturas. Denomina-se α, o autovalor com maior magnitude, e β, o de menor. Pode-se, então, calcular a soma dos autovalores pelo traço de $\mathbf{H}$ e o produto pelo determinante:

$$Tr(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta \tag{2.27}$$

$$Det(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta \tag{2.28}$$

Para o caso em que o determinante é negativo, as curvaturas possuem sinais diferentes e então o ponto é descartado como não sendo um extremo. Sendo r a razão entre o autovalor de maior magnitude e o de menor, de modo que $\alpha = r\beta$, temos:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} = \frac{(\alpha+\beta)^2}{\alpha\beta} = \frac{(r\beta+\beta)^2}{r\beta^2} = \frac{(r+1)^2}{r} \tag{2.29}$$

A Equação 2.29 depende apenas da razão entre os autovalores, independente de seus valores individuais. O valor $(r+1)^2/r$ é mínimo quando os dois autovalores são idênticos e cresce com r. Portanto, para conferir se a razão entre as curvaturas está abaixo de um determinado limiar r, basta avaliar a Equação 2.30:

$$\frac{Tr(\mathbf{H})^2}{Det(\mathbf{H})} < \frac{(r+1)^2}{r} \tag{2.30}$$

Lowe propõe o uso de r=10.

Após a seleção dos pontos de interesse, a próxima etapa definirá a orientação para cada um deles.

3. **Definição de orientação:** Nesta etapa define-se a orientação dos pontos de interesse selecionados. Ao se atribuir uma orientação consistente para cada ponto de interesse, podem-se representar os descritores em relação a esta orientação, conseguindo-se assim, invariância quanto à rotação. O método utilizado para se atribuir esta orientação é apresentado como se segue.

   Para determinar a orientação dos pontos de interesse, é calculada a magnitude e a orientação do gradiente para cada *pixel* da imagem. Assim, é gerado o histograma das orientações do gradiente dos *pixels* em uma região vizinha do ponto de interesse, sendo que picos nesse histograma correspondem a direções dominantes para os gradientes locais. O maior pico do histograma e outros picos que correspondam a no mínimo 80% do valor do maior pico serão usados para criar um ponto de interesse com aquela orientação, ou seja, para locais de múltiplos picos de magnitude semelhante, são criados diferentes pontos de interesse na mesma localização, mas com diferentes orientações. Finalmente, uma parábola é usada para interpolar os três valores do histograma mais próximos do pico, de forma a se ter uma melhor precisão de sua posição.

4. **Descrição dos pontos de interesse:** Nas etapas anteriores, para cada oitava, foram selecionados pontos de interesse para localizações, escala σ e orientações definidos. Nesta última etapa, é apresentada a criação do descritor que representa as regiões relativas a cada um destes pontos. Este descritor é criado computando-se as magnitudes e orientações dos gradientes que são amostrados ao redor da localização

do ponto, na escala selecionada. Para alcançar a invariância à orientação, as coordenadas do descritor são rotacionadas em relação à orientação do ponto de interesse, calculada na etapa anterior. A Figura 2.24 mostra a área ao redor de um ponto de interesse (imagem superior esquerda) a ser considerada após realizar a rotação da coordenada do descritor em relação a orientação do ponto.

A Figura 2.24 apresenta o esquema de geração do vetor de características para um determinado ponto de interesse. A janela (quadriculado verde da imagem superior direita) é definida em n x n regiões, com k x k *pixels* cada, ao redor da localização do ponto de interesse. Lowe propõe a utilização do valor quatro para n e k. Assim, as magnitudes do gradiente e orientações são amostradas em volta da localização do ponto de interesse, usando a escala do ponto (como representado pelas setas pequenas em cada localização amostrada na imagem superior direita). Além disso, uma função Gaussiana com σ igual à metade da largura da janela do descritor é usada para associar um peso à magnitude do gradiente de cada ponto amostrado (como ilustrado na imagem superior direita com a janela circular). O objetivo dessa janela Gaussiana é evitar mudanças repentinas no descritor com pequenas mudanças na posição da janela, e dar menos ênfase aos gradientes que estão distantes do centro do descritor.

Após essa suavização dos gradientes, são criados histogramas para oito direções em cada região n mostrada na Figura 2.24 (imagem inferior direita). O histograma é criado com as magnitudes dos *pixels* pertencentes a cada região. O comprimento de cada seta corresponde à soma das magnitudes dos gradientes perto dessa direção dentro da região. Por fim, o descritor é formado por um vetor contendo as magnitudes de todas as orientações dos histogramas correspondentes aos tamanhos dessas setas mostradas conforme apresentado no esquema. Temos então um vetor de características de dimensionalidade 128 (4 x 4 histogramas de 8 orientações cada um).

Para que o descritor gerado tenha invariância à iluminação, este é normalizado. Após a normalização, todos os valores acima de um determinado limiar são ajustados para este limiar. Isto é feito para que direções com magnitude muito grande não dominem a representação do descritor.

Por fim, ao finalizar as 4 etapas de aplicação do descritor SIFT em uma determinada imagem, serão obtidos vários vetores de características de tamanho 128, um para cada ponto de interesse. A Figura 2.25 mostra o resultado do SIFT aplicado a uma determinada imagem mostrando os pontos de interesse detectados juntamente com suas magnitudes e orientações.

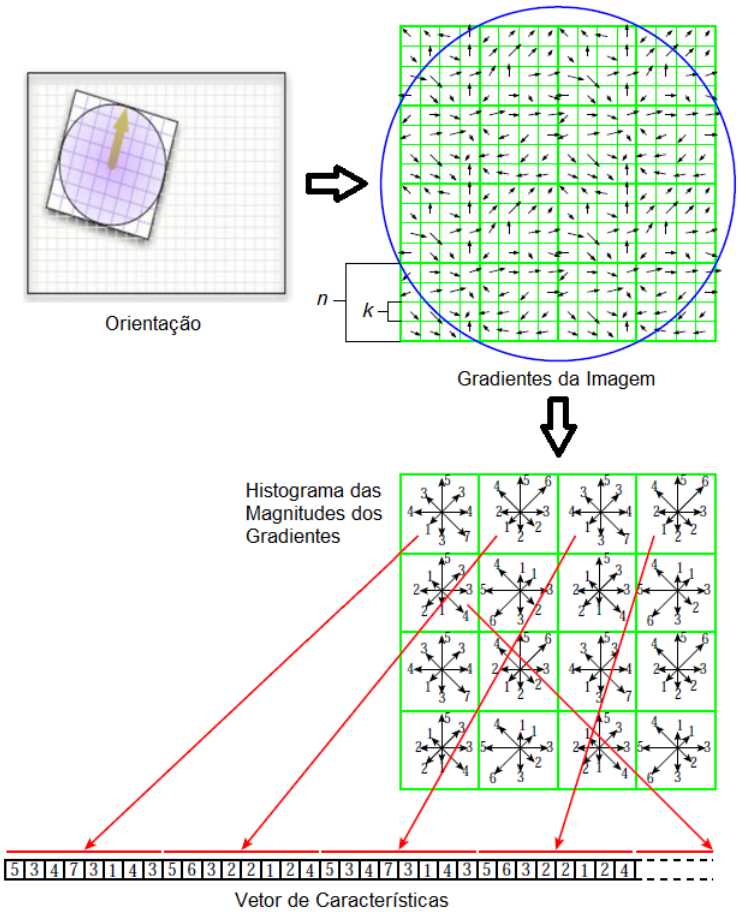

Figura 2.24: Esquema de criação do vetor de característica para um determinado ponto de interesse. n corresponde a uma região e k a um *pixel*.

2.3.2 Construção do Dicionário de Palavras Visuais

Uma vez extraídas todas as características das regiões associadas aos pontos de interesse é criado o dicionário de palavras visuais. Uma das maneiras mais comuns, apresentada na literatura, para se criar o dicionário é aplicando o método de agrupamento *k-means* [Jain et al. 1999] ao conjunto de todos os descritores SIFT extraídos da imagem conforme [Sivic e Zisserman 2003], [Csurka et al. 2004], [Winn et al. 2005], [Wang et al. 2008], [Batista et al. 2009], [Lopes et al. 2009b], [Lopes et al. 2009a].

O método de agrupamento *k-means* tem como objetivo gerar k grupos disjuntos de forma a minimizar a distância intra-classe e maximizar a distância inter-classe. Para cada grupo gerado, tem-se o representante do mesmo, chamado de centróide. Cada centróide representa uma palavra visual do dicionário de palavras visuais. Dessa forma, o valor de k representa a quantidade de grupos gerados pelo *k-means* e consequentemente o tamanho do dicionário de palavras visuais.

O tamanho do dicionário deve ser grande o suficiente para distinguir mudanças relevantes nas partes da imagem, mas não tão grande a ponto de distinguir variações irrelevantes como ruídos [Csurka et al. 2004]. Para este método de agrupamento, conforme mostra a

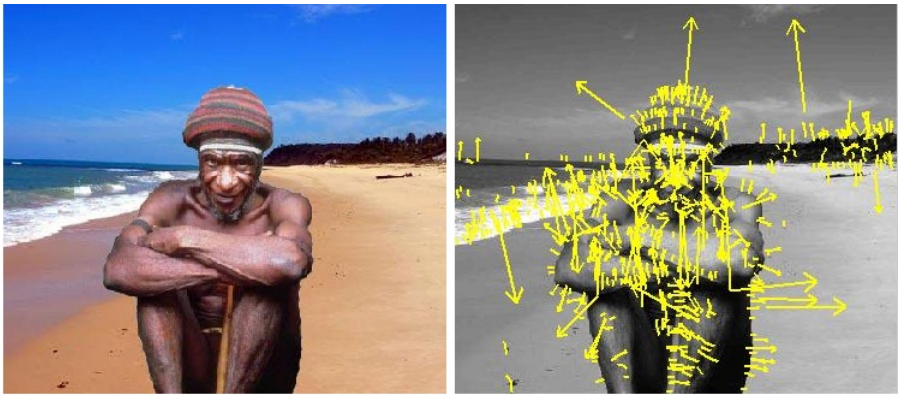

Figura 2.25: Imagem apresentando os pontos de interesse, a magnitude e a orientação detectados pelo SIFT.

literatura, é comum escolher o valor de k de forma empírica.

A Figura 2.26 ilustra o processo de criação de um dicionário de palavras visuais contendo três palavras visuais. Inicialmente são extraídos todos os descritores SIFT (representadas pelos círculos pretos) de todas as imagens da base de dados. Esse conjunto de descritores são agrupados pelo *k-means* com k=3 e o resultado é a separação de três grupos de cores diferentes (vermelho, amarelo e verde) cujos elementos são representados por círculos e o centróide por quadrados. Os três centróides gerados compõem o dicionário de palavras visuais.

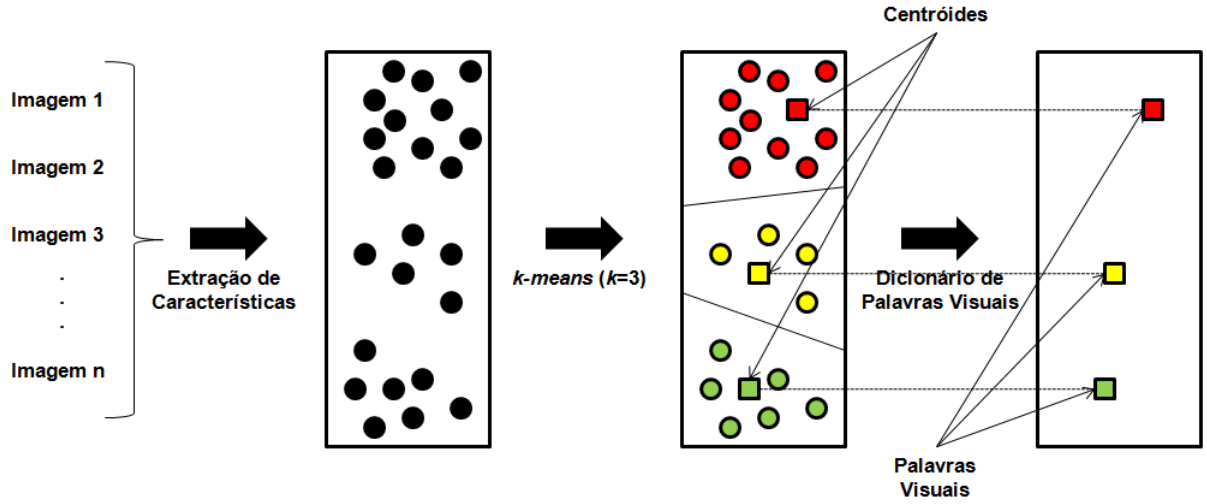

Figura 2.26: Ilustração do processo de criação do dicionário de palavras visuais.

2.3.3 Construção dos Histogramas de Palavras Visuais

Nesta fase é criado um histograma das palavras visuais para cada imagem da base de dados. Este histograma possui o tamanho do dicionário de palavras visuais e cada posição representa a frequência com que cada palavra visual ocorre na imagem.

Para definir as frequências do histograma, são calculadas as distâncias de cada vetor de características da imagem a todos os centróides (palavras visuais) determinados pelo k-

means para um determinado k escolhido. As distâncias são calculadas utilizando alguma métrica de similaridade, como por exemplo a distância Euclidiana. Cada vetor de característica da imagem é associado à palavra visual que possuir a menor distância do vetor. Assim, no cálculo da frequência das palavras visuais soma-se a quantidade de vetores que estão associados a palavra um, depois a palavra dois até a palavra k, que corresponde a última palavra do dicionário. Finalmente temos o histograma com k frequências de cada palavra visual do dicionário que ocorre na imagem. Este histograma de palavras visuais resultante compõe as características que serão utilizadas no processo de recuperação da imagem.

2.3.4 Busca por Similaridade

A última fase da abordagem *bag-of-visual-features* corresponde à busca por similaridade das imagens utilizando os histogramas de frequência gerados na fase anterior. Dada uma imagem de consulta, um determinado operador de similaridade e uma métrica de similaridade (nesta fase é mais comum a utilização da distância Cosseno), são recuperadas todas as imagens similares à imagem de consulta. A similaridade é obtida a partir do cálculo da distância do histograma da imagem de consulta em relação a todos os histogramas das imagens da base de dados. Quanto menor a distância, mais similar.

2.4 Atenção Visual

O sistema visual dos vertebrados superiores possui uma habilidade chamada de atenção visual que é responsável por selecionar e processar rapidamente somente as regiões mais relevantes em uma determinada cena visual. Esta seleção das informações mais relevantes que estimulam o campo visual é uma das características mais importantes dos sistemas visuais biológicos que permite rápida detecção de predadores, perpetuação e evolução das espécies [Itti e Koch 2001]. Além disso, devido ao grande volume de informações que estimulam o campo visual em sistemas biológicos, a seleção de informações relevantes de uma cena passa a ser muito importante, pois reduz a quantidade de informações a serem processadas [Fischer e Weber 1993]. No processamento de imagens, em sistemas computacionais, tem-se o mesmo problema do grande volume de informações a serem processadas. Assim, a atenção visual passa a ser uma alternativa para esse problema.

Em sistemas computacionais, a atenção visual pode ser obtida basicamente por dois tipos de métodos, os métodos *bottom-up* e os métodos *top-down*. Os métodos *bottom-up* consideram características de baixo nível das imagens (cor, intensidade e orientação), sem qualquer informação contextual, para definir a atenção visual. Já os métodos *top-down* utilizam as características de alto nível das imagens (conhecimento prévio, modelos geométricos, modelos estatísticos, entre outros) para detectar as regiões de maior interesse

da cena.

Nas próximas seções são apresentados dois modelos *bottom-up* de extração da atenção visual de uma determinada imagem, o modelo de Itti [Itti et al. 1998] e o modelo de Harel [Harel et al. 2007]. Ambos os modelos foram utilizados neste trabalho.

2.4.1 Modelo de Itti

Uma das técnicas mais encontradas na literatura para extrair a atenção visual de imagens é o modelo de Itti [Itti et al. 1998] que se baseia na extração de mapas de saliência. Este método decompõe a imagem em um conjunto de mapas de características topográficas. Em seguida, diferentes regiões espaciais competem entre si pela saliência de cada mapa, porém apenas regiões que localmente se destacam conseguem persistir. Por fim, todos os mapas de características gerados alimentam, de maneira *bottom-up*, um mapa de saliência final, que codifica topograficamente a conspicuidade ou regiões salientes locais de toda a cena. Para isso, o modelo de Itti é dividido nos seguintes processos: filtragem linear, diferenças centro-vizinhanças e normalização, combinações variando a escala e normalização, combinações lineares e rede neural. A Figura 2.27 apresenta a arquitetura *bottom-up* do modelo de Itti.

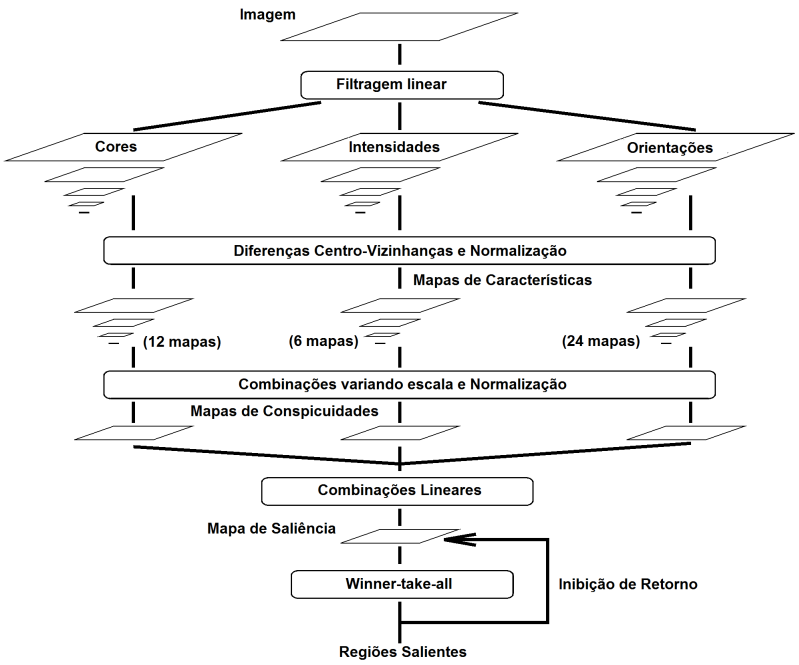

Figura 2.27: Arquitetura do modelo de Itti. Ilustração baseada no artigo [Itti et al. 1998].

Para gerar um mapa de saliência, três tipos de características visuais primitivas são extraídas: cor, intensidade e orientação. Em seguida, quatro canais de cores são criados (R para vermelho, G para verde, B para azul e Y para amarelo). Sendo r, g, b os canais

vermelho, verde e azul da imagem de entrada, os canais de cores são representados por:

$$R = r - (g + b)/2 \tag{2.31}$$

$$G = g - (r + b)/2 \tag{2.32}$$

$$B = b - (r + g)/2 \tag{2.33}$$

$$Y = (r + g)/2 - |r - g|/2 - b \tag{2.34}$$

A imagem de intensidade é representada por I=$(r$+g+$b)/3$, que define a imagem em tons de cinza. Tanto I, como R, G, B e Y são utilizados para criarem Pirâmides Gaussianas [Greenspan et al. 1994] *I(σ)*, *R(σ)*, *G(σ)*, *B(σ)* e *Y(σ)* em que $\sigma \in \{0..8\}$ é o fator de escala. Portanto, nove escalas espaciais são criadas para I, R, G, B e Y. A Pirâmide Gaussiana é composta por versões suavizadas usando filtros Gaussianos aplicados à imagem de entrada.

Os mapas de características são obtidos por meio da diferença entre canais de cores em diferentes escalas, este processo é conhecido como diferença centro-vizinhança, sendo definido por "$\ominus$". O centro é um *pixel* na escala $c \in \{2,3,4\}$, e sua vizinhança corresponde ao *pixel* na escala s=c+δ, em que $\delta \in \{3,4\}$.

Dessa forma, é gerado um conjunto de 6 mapas de características de intensidade $\mathcal{I}(c,s)$ com $c \in \{2,3,4\}$ e s=c+δ, onde $\delta \in \{3,4\}$:

$$\mathcal{I}(c, s) = |I(c) \ominus I(s)| \tag{2.35}$$

Um segundo conjunto de 12 mapas de características são gerados para cor. As Equações 2.36 e 2.37 definem matematicamente as diferenças centro-vizinhanças:

$$\mathcal{RG}(c, s) = |(R(c) - G(c)) \ominus (G(s) - S(s))| \tag{2.36}$$

$$\mathcal{BY}(c, s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \tag{2.37}$$

Por fim, informação de orientação local é obtida de I utilizando Pirâmides de Gabor $O(\sigma, \theta)$, em que $\sigma \in \{0..8\}$ representa a escala e $\theta \in \{0°,45°,90°,135°\}$ é a orientação preferencial. Vinte e quatro mapas de características são extraídos pela diferenças centro-vizinhanças conforme Equação 2.38:

$$\mathcal{O}(c, s, \theta) = |O(c, \theta) \ominus O(s, \theta)| \tag{2.38}$$

Os mapas de características gerados, são por sua vez combinados em três mapas de conspicuidades, $\overline{\mathcal{I}}$ para intensidade (Equação 2.39), $\overline{\mathcal{C}}$ para cor (Equação 2.40) e $\overline{\mathcal{O}}$ para orientação (Equação 2.41), na escala σ=4. A motivação para a criação de três canais separados ($\overline{\mathcal{I}}$, $\overline{\mathcal{C}}$ e $\overline{\mathcal{O}}$) é a hipótese de que características similares competem pela saliência, enquanto que características diferentes contribuem independentemente para o mapa de saliência.

O propósito do mapa de saliência é representar as conspicuidades ou regiões salientes na imagem por quantidades escalares e guiar a seleção de regiões baseada na distribuição espacial da saliência. As Equações 2.39 a 2.41 modelam matematicamente a obtenção dos mapas de conspicuidades, em que $\mathcal{N}$ representa a normalização e o operador "$\bigoplus$"representa a adição em escala que consiste na redução de cada mapa a escala quatro:

$$\overline{\mathcal{I}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c=4} \mathcal{N}(\mathcal{I}(c,s)) \tag{2.39}$$

$$\overline{\mathcal{C}} = \bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} [\mathcal{N}(\mathcal{RG}(c,s)) + \mathcal{N}(\mathcal{BY}(c,s))] \tag{2.40}$$

$$\overline{\mathcal{O}} = \sum_{\theta \in \{0°,45°,90°,135°\}} \mathcal{N}(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{N}(\mathcal{O}(c,s,\theta))) \tag{2.41}$$

Para criar o Mapa de Saliência, os 3 mapas de conspicuidades são normalizados e somados conforme Equação 2.42.

$$S = \frac{1}{3}(\mathcal{N}(\overline{\mathcal{I}}) + \mathcal{N}(\overline{\mathcal{C}}) + \mathcal{N}(\overline{\mathcal{O}})) \tag{2.42}$$

O Mapa de Saliência resultante é uma imagem em tons de cinza em que as regiões mais salientes são representadas por *pixels* de altas intensidades. Desta forma, podem ocorrer regiões que possuem *pixels* com valores iguais. Para evitar que uma mesma região seja determinada como mais saliente mais de uma vez e para que seja possível determinar várias regiões salientes, mesmo que tais regiões possuam *pixels* de mesmo valor, utiliza-se o princípio da inibição de retorno segundo o qual o Mapa de Saliência gerado alimenta uma rede neural *Winner-Take-All* (WTA) [Koch e Ullman 1985] na escala σ=4, que por sua vez faz com que suas interações sinápticas garantam a manutenção das regiões mais importantes, enquanto que as outras regiões são inibidas. A Figura 2.28 mostra um exemplo de uma imagem e seu respectivo Mapa de Saliência.

2.4.2 Modelo de Harel

Harel et al. usaram um outro modelo *bottom-up* de extração de atenção visual chamado *Graph-Based Visual Saliency* (GBVS) [Harel et al. 2007]. Baseando-se na arquitetura de

Figura 2.28: Mapa de Saliência extraído utilizando o modelo de Itti.

Itti [Itti et al. 1998], explicada na subseção anterior 2.4.1, foi proposto uma alternativa para as fases de diferenças centro-vizinhanças/normalização e combinações variando a escala/normalização conforme mostrado na Figura 2.27. Essas novas fases foram chamadas de ativação e normalização/combinação (mostradas na Figura 2.29 em azul) respectivamente.

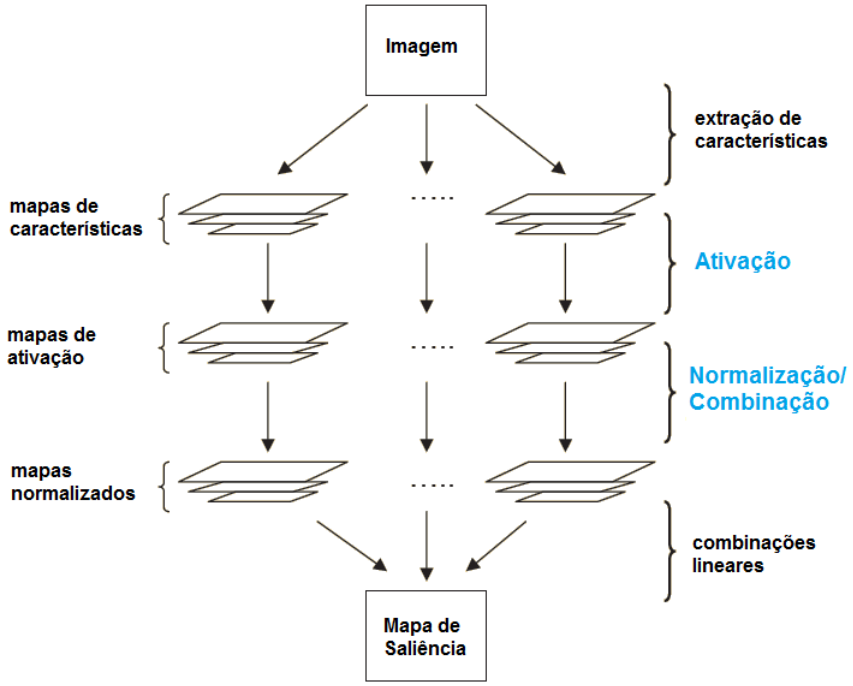

Figura 2.29: Arquitetura do modelo de Itti. Ilustração baseada no artigo [Harel et al. 2007].

Na biologia, o córtex visual possui neurônios conectados em uma rede comunicando-se uns com os outros através de sinápsis permitindo a visualização e análise rápida das regiões mais importantes de uma determinada cena [Harel et al. 2007]. Dessa forma, o modelo de Harel propõe uma solução baseada em grafos que processa as regiões locais levando em consideração as informações globais para obter um mapa de saliência. Para os dois processos citados, ativação e normalização/combinação, são construídos grafos direcionais com pesos nas arestas a partir dos mapas. Os grafos são tratados como uma

Cadeia de Markov para calcular a distribuição de equilíbrio.

Na fase de Ativação proposta, têm-se como entrada os mapas de características gerados pela fase de extração de características, conforme mostra a Figura 2.29. Os mapas são representados como $M:[n]^2 \rightarrow \mathbb{R}$. O objetivo dessa fase de Ativação é gerar o mapa de ativação $A:[n]^2 \rightarrow \mathbb{R}$, tal que os valores de $M(i,j)$ que forem incomuns em sua vizinhança corresponderão a valores elevados em A.

Dessa forma para se obter A, é criado um grafo direcionado G_A totalmente conectado. O grafo é montado da seguinte forma:

1) Um nó é instanciado em todas as regiões do mapa (características ou ativação) de entrada conforme mostra a Figura 2.30:

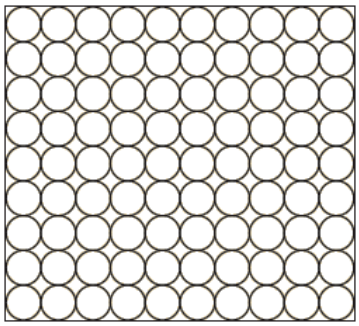

Figura 2.30: Representação dos nós em um mapa (características ou ativação) de entrada.

2) Criam-se arestas bidirecionais em todos os nós do grafo tornando-o totalmente conectado, conforme mostra a Figura 2.31:

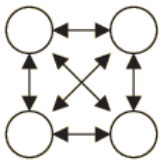

Figura 2.31: Representação das arestas bidirecionais do grafo.

3) Ponderam-se as arestas, conforme mostra a Figura 2.32:

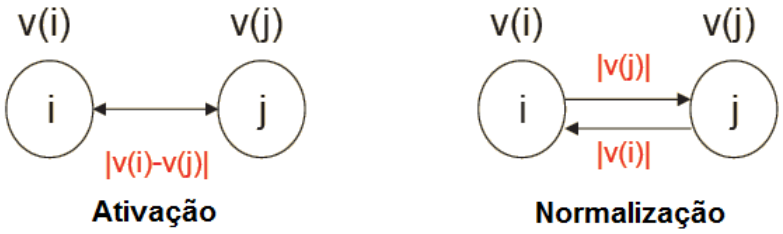

Figura 2.32: Representação da ponderação das arestas no grafo direcionado tanto para a fase de Ativação quanto para a fase de Normalização.

Conforme mostrado no item 3, as arestas dos nós são ponderadas. Assim, a aresta direcionada do nó (i,j) para o nó (p,q) terá sua ponderação calculada conforme as Equações de 2.43 a 2.45:

$$w_1((i,j),(p,q)) \triangleq d((i,j)||(p,q))F(i-p,j-q), emque \tag{2.43}$$

$$d((i,j)||(p,q)) \triangleq |\log \frac{M(i,j)}{M(p,q)}| \tag{2.44}$$

$$F(a,b) \triangleq \exp(-\frac{a^2+b^2}{2\sigma^2}) \tag{2.45}$$

Uma vez calculada as ponderações para G_A, o cálculo da distribuição de massa é realizado em um ambiente síncrono, em que a cada passo, cada nó soma as massas de entrada e em seguida passam-se as medidas dessa massa para seus vizinhos de acordo com os pesos calculados nas arestas. Este mesmo processo acontecendo nos nós simultaneamente dá origem a uma distribuição equilibrada de massa, formando então o mapa de ativação A.

A próxima fase proposta é a de Normalização, em que o mapa de ativação A é normalizado com o objetivo de concentrar suas massas. Considerando que $A{:}[n]^2{\rightarrow}\mathbb{R}$ é construído o grafo direcionado G_N com n^2 nós. Para cada nó (i,j) e todos os nós (p,q) (incluindo (i,j)) ao qual é conectado, é gerada a aresta de (i,j) para (p,q) com o peso calculado pela Equação 2.46:

$$w_2((i,j),(p,q)) \triangleq A(p,q)F(i-p,j-q) \tag{2.46}$$

Assim, normalizando os pesos das arestas de saída de cada nó e tratando o resultado do grafo como uma cadeia de Markov nos permite calcular a distribuição de equilíbrio sobre todos os nós. A massa acumula preferencialmente nos nós com alta ativação. A Figura 2.33 mostra um exemplo de uma imagem e seu respectivo Mapa de Saliência.

Figura 2.33: Mapa de Saliência extraído utilizando o modelo de Harel.

2.5 Considerações Finais

Neste Capítulo foram apresentados os fundamentos teóricos necessários para a elaboração deste trabalho. Inicialmente foi descrito o processo de recuperação de imagens por conteúdo detalhando as principais fases do seu fluxo de funcionamento: extração de características, medidas de similaridade e operadores de similaridade. Durante a apresentação da fase de extração de características, foram mostrados alguns descritores comuns na literatura e que foram utilizados como parâmetro de comparação nos experimentos deste trabalho.

Na seção seguinte, com o intuito de demonstrar as métricas de avaliação dos experimentos realizados, foram apresentadas as medidas de avaliação utilizadas em sistemas de recuperação. Medidas como Precisão, Revocação, *MAP* e *NDCG*.

Como este trabalho é uma extensão da abordagem *bag-of-visual-features*, foi elaborada uma seção para descrever toda essa abordagem exemplificando-a com alguns trabalhos encontrados na literatura. As fases utilizadas para descrever a abordagem foram: extração de características, construção do dicionário de palavras visuais, construção dos histogramas de palavras visuais e busca por similaridade. Toda essa abordagem foi implementada no trabalho modificando a fase de construção dos histogramas de palavras visuais.

Por fim, a última seção descreve a importância de técnicas de extração da Atenção Visual para os sistemas computacionais relacionado principalmente ao processamento de imagens. Nesta seção são apresentados dois modelos de extração de mapas de saliências de imagens, o modelo de [Itti et al. 1998] e o modelo de [Harel et al. 2007], ambos utilizados neste trabalho para considerar a percepção visual humana durante a construção dos histogramas de palavras visuais.

Capítulo 3

Estado da Arte

Neste capítulo são descritos os principais trabalhos encontrados na literatura relacionados à classificação e recuperação de imagens considerando à percepção visual humana.

3.1 Trabalhos Relacionados

O reconhecimento de objetos é uma questão muito importante na visão computacional especialmente na presença de diferentes condições de iluminação, cor, escala, translação de objetos, mudanças de fundo da imagem, entre outros. As imagens com essas variações são conhecidas como *Near Duplicate Images*, a Figura 3.1 mostra alguns exemplos dessas imagens que diferem umas das outras em algum tipo de transformação. *Near Duplicate Retrieval*, a técnica responsável pela recuperação de imagens *Near Duplicate*, é especialmente importante na detecção da utilização ilegal de imagens / vídeo ou violação de direitos autorais, na recuperação de sub-imagens e na filtragem de imagens *spam* [Yang et al. 2009] [Zhang e Chang 2004].

Vários trabalhos têm sido propostos com o objetivo de reconhecimento de objetos. Métodos globais ou locais com base na cor, forma ou textura são sensíveis às variações de imagem, tais como aquelas ilustradas na Figura 3.1. O *bag-of-visual-features*, utilizando descritores SIFT, é uma técnica muito utilizada na recuperação e/ou classificação de imagens e se mostra robusta em mudanças de cor, iluminação e transformações de escala, no entanto, a informação espacial das palavras visuais é negligenciada nessa técnica. São diversos os estudos nessa área visando tornar este processo cada vez mais genérico, independente do domínio das imagens, levando em consideração a eficiência computacional e a qualidade da recuperação e/ou classificação das imagens. Para isso, vários aspectos dessa abordagem são analisados, como por exemplo, a diminuição de palavras visuais do dicionário mantendo a sua qualidade discriminativa, a escolha do extrator de características para montar os vetores de características, a escolha do processo de quantização do dicionário, a melhoria no processo de indexação das palavras visuais, entre outros.

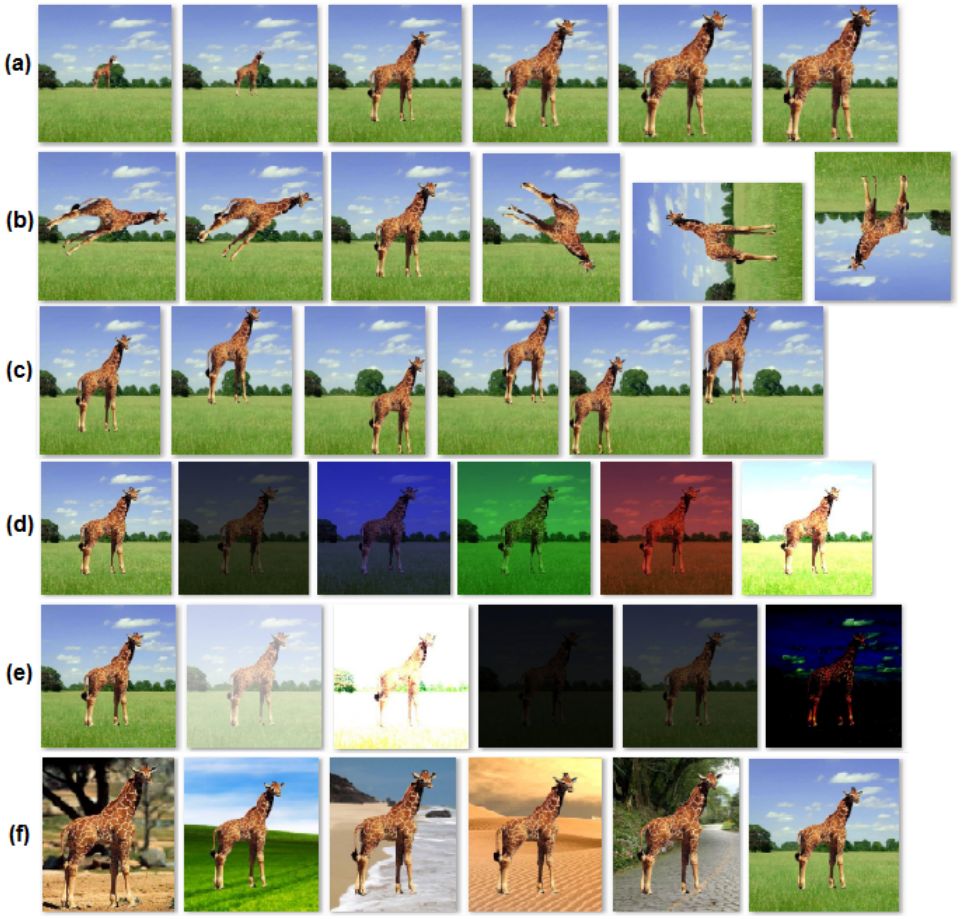

Figura 3.1: Figure 1. Exemplo de imagens *Near Duplicate* em diversas variações: (a) escala do objeto da cena, (b) rotação do objeto e da imagem, (c) translação do objeto, (d) mudança de cor da imagem, (e) mudança da iluminação/contraste da imagem, (f) mudança do fundo da imagem.

Apesar de existirem vários esforços com o intuito de aperfeiçoar a abordagem *bag-of-visual-features*, uma melhoria tem sido pouco explorada, a qual sugere o processamento prévio das imagens separando as partes relevantes das partes irrelevantes a serem recuperadas e/ou classificadas. Esse processamento ocorre de forma automática para a percepção humana pois ao analisar a Figura 3.2, por exemplo, é perceptível que o foco principal da imagem seja o cavalo. No entanto, ao se extrair as características dessa imagem, misturam-se características do cavalo com características do fundo da cena, prejudicando a recuperação e/ou classificação.

Um dos primeiros trabalhos a considerar a utilização da percepção humana foi proposto por [Moosmann et al. 2006]. Neste trabalho, é aplicada a técnica de extração de mapas de saliência proposta por [Itti et al. 1998] com o objetivo de classificar objetos em situações adversas. Os autores propõem um sistema de classificação, no qual os mapas de saliência são montados durante o processo de classificação da imagem. O processo funciona da seguinte maneira: inicialmente, janelas limitantes, com tamanhos aleatórios, são amostradas, também aleatoriamente, nas imagens de treinamento para

Figura 3.2: Imagem que tem como foco principal o cavalo na cena.

extrair as características das mesmas. Árvores de decisão aleatórias (*EXTremely RAndomized Trees* [Marée et al. 2005]) são construídas a partir das características extraídas dessas janelas como um classificador. Durante a fase de testes, as janelas são novamente amostradas aleatoriamente nas imagens de testes e classificadas pelas árvores de decisão. No momento da classificação o classificador é utilizado para montar o mapa de saliência da imagem analisada, ajudando a identificar a janela amostrada como parte do objeto ou parte do fundo da imagem.

O mapa de saliência descrito é definido como uma função de densidade de probabilidade tridimensional, no qual o ponto da função é definido como saliente se o mesmo for classificado como não integrante do fundo (*background*) da imagem. Para isso o autor considera as regiões de alta saliência dos mapas como sendo as regiões representantes do *foreground* e as regiões de baixa saliência como sendo as regiões representantes de *background* da imagem. Foram realizados experimentos em três bases de dados: GRAZ-02 [Opelt e Pinz 2005], Pascal 2005 [Everingham et al. 2005] e uma base de dados de imagens de cavalos utilizada em [Jurie e Schmid 2004]. Os resultados mostraram ganhos na utilização da técnica de extração de mapas de saliência comparada com a não utilização, porém o autor propõe mais testes em bases de dados com mais classes de imagens.

Ainda no ano de 2006, [Marszalek e Schmid 2006] propuseram uma extensão do processo de classificação de categorias utilizando a abordagem *bag-of-visual-features*. O método tem como objetivo explorar a relação espacial entre as características das imagens ponderando-as, dando mais importância às características do objeto de interesse (*foreground*) e menos para às características de fundo da imagem (*background*). Para isso, propuseram um método que produz uma máscara de segmentação baseada na informação das imagens de treinamento. Essa máscara é utilizada para ponderar as características de *foreground* e *background*, dando mais importância ao *foreground*.

Durante a fase de treinamento do classificador, são utilizadas imagens de treinamento que possuem o *foreground* marcado manualmente representando a segmentação ideal da imagem (*ground-truth*) de acordo com a aplicação dos autores. A Figura 3.3 mostra exemplos da anotação dos objetos de interesse das imagens fornecidas pela base PASCAL [Everingham et al. 2005]. Uma vez preparado o conjunto de treinamento, são gera-

das hipóteses sobre a possível localização de objetos e formas para cada característica das imagens de testes. Com base nessas hipóteses são montadas as máscaras de segmentação para as imagens de teste, as quais são consideradas como mapas de pontuação descrevendo a probabilidade de um determinado *pixel* da imagem pertencer a um objeto de interesse. Por fim, na construção do histograma das imagens de testes, as características são ponderadas de acordo com o valor correspondente na máscara.

Figura 3.3: Exemplos de imagens da base PASCAL [Everingham et al. 2005] anotadas manualmente. Ilustração retirada do artigo [Marszalek e Schmid 2006].

Os experimentos mostraram uma melhora na classificação utilizando a técnica proposta por Marszaleck, no entanto vale ressaltar que apesar da máscara ser semelhante aos mapas de saliência, seu processo de geração é diferente e dependente de um conjunto de imagens de treinamento previamente segmentadas. Além disso, a técnica leva em consideração a classificação de objetos e portanto a maior importância sempre é dada ao *foreground* da imagem não permitindo, por exemplo, classificar uma determinada cena, ou seja, permitir dar uma maior importância às características de *background*.

O trabalho de [Huang et al. 2008] propõe um novo método de representação de imagens que trata as imagens como histogramas de características salientes. Para isso, considerando a abordagem *bag-of-visual-features*, ao invés de se construir os histogramas de palavras visuais baseando-se no dicionário de palavras visuais, o autor propõe a construção dos histogramas de frequência por meio da melhor projeção do subespaço de características extraídas das imagens. Essa projeção é construída pela redução de dimensionalidade do universo de características utilizando a técnica de análise discriminante linear (*linear discriminant analysis - LDA*) [Borg e Groenen 2005] que se propõe a minimizar a variância intra-classes e a maximizar a variância inter-classes.

Antes de realizar a projeção do subespaço de características, são extraídos os mapas de saliência das imagens. Os autores propuseram uma nova estratégia de extração dos mapas de saliência que enfatiza as regiões que contêm bordas e cantos diferentemente das

que se baseiam em cor, intensidade e orientação da informação. A Figura 3.4 apresenta os mapas de saliência e a segmentação binária desses mapas para a nova estratégia e para a técnica de [Itti et al. 1998]. Utilizou-se a técnica apresentada em [Otsu 1979] para encontrar o *threshold* ótimo das segmentações.

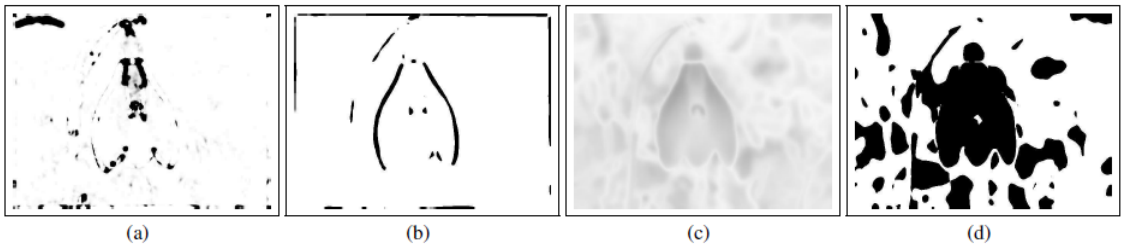

Figura 3.4: Comparação da estratégia de extração de mapas de saliência: (a) mapa de saliência extraído utilizando a estratégia de [Huang et al. 2008], (b) segmentação *foreground/backgorund* da imagem (a), (c) mapa de saliência extraído utilizando a estratégia de [Itti et al. 1998], (b) segmentação *foreground/backgorund* da imagem (c). Ilustração retirada do artigo [Huang et al. 2008].

Nos experimentos apresentados em [Huang et al. 2008] é utilizada uma base de dados de imagens de flores, chamada Oxford [Nilsback e Zisserman 2006]. O método proposto de representação das imagens por histogramas de características salientes foi comparado a estratégia de representação por histogramas de palavras visuais utilizando o dicionário de palavras como apresentado em [Nilsback e Zisserman 2006]. Além disso, para avaliar também a estratégia de extração de mapas saliências utilizada para criar os histogramas de características salientes, comparou-se a mesma com a técnica de extração de mapas de saliência proposta por [Itti et al. 1998]. Os resultados mostraram que as técnicas propostas obtiveram melhores resultados na recuperação KNNQ caso k fosse maior ou igual a 3. Além disso, verificou-se que a técnica proposta de extração dos mapas saliência era mais eficiente computacionalmente.

[Sato e Katto 2010] desenvolveram um trabalho no qual propuseram a remoção de áreas irrelevantes nas imagens para melhorar a fase de aprendizagem em um processo de classificação. Para isso, em conjunto com a abordagem de *bag-of-visual-features*, utilizaram a técnica de extração dos mapas de saliência proposto por [Itti et al. 1998] e um método chamado *Seam Carving* [Avidan e Shamir 2007] responsável por redimensionar uma imagem removendo conjuntos de pixels com baixas energias. Inicialmente a imagem é processada gerando o seu mapa de saliência. A técnica *Seam Carving* é aplicada na imagem e, por meio de uma função de energia da mesma, reduz-se a imagem removendo algumas regiões desnecessárias, conforme mostrado na Figura 3.5. Por fim, na geração dos vetores de características, são considerados apenas os pontos-chave que fazem parte da área da imagem de maior interesse definida pelo mapa de saliência. Não é apresentada nenhuma estratégia de selecionar ou ponderar os pontos-chave usando mapa de saliência. Nos experimentos foi utilizada a base de dados Caltech-101 [Fei-Fei et al. 2007]. Os re-

sultados obtidos foram melhores comparados com a abordagem de *bag-of-visual-features* sem utilizar mapas de saliência.

Figura 3.5: (a) imagem original (b) imagem após aplicar a técnica *Seam Carving*. Ilustração retirada do artigo [Sato e Katto 2010].

Um outro trabalho relacionado ao tema consiste em criar um novo descritor de características chamado *Salient*-SIFT. Este trabalho foi proposto por Liang [Liang et al. 2010] e tem como objetivo extrair e descrever as características das imagens considerando as regiões de interesse de acordo com a percepção visual humana.

Inicialmente é utilizada a técnica de extração de mapas de saliência proposta por [Itti et al. 1998]. Após isso, é aplicado nos mapas de saliência o detector de bordas proposto por [Canny 1986] para que a região de interesse da imagem seja selecionada. Utilizando a Triangulação de Delaunay [de Berg et al. 2008], é gerado um contorno do objeto da imagem. Esse contorno é tratado como a região de saliência que será utilizada para selecionar os pontos-chave da imagem a serem descritos pelo SIFT. A Figura 3.6 mostra exemplos dos descritores montados pelo processo apresentando a imagem original, o contorno gerado após aplicar o método proposto, os descritores gerados pelo SIFT em toda a imagem e os descritores gerados pelo SIFT apenas nos pontos contidos no contorno.

Essa abordagem diferencia um pouco das demais principalmente porque o autor incluiu o processo de seleção da região de interesse dentro do descritor de características. Essa estratégia não permite ponderar o *background* e o *foreground* de acordo com a necessidade de recuperação. Os experimentos foram realizados na base Caltech-256 [Griffin et al. 2007] e segundo os resultados do autor, o qual apenas comparou o descritor SIFT com o descritor *Salient*-SIFT, este último teve melhores resultados.

A percepção visual humana também é abordada no trabalho de Shogo [Nakamoto e Toriu 2011]. Neste trabalho é investigada a eficácia dos mapas de saliência para reduzir os pontos-chave extraídos das imagens considerando somente aqueles que melhor discriminam o objeto de interesse da imagem. Para isso o autor utiliza um *threshold* nos mapas de saliência variando o seu percentual em 0, 25, 50, 75 e 100 (a Figura 3.7 mostra a seleção dos pontos-chave variando o percentual de *threshold*) para eliminar os pontos-chave de menor importância. Foram realizados testes em duas bases de dados, Caltech-256 [Griffin

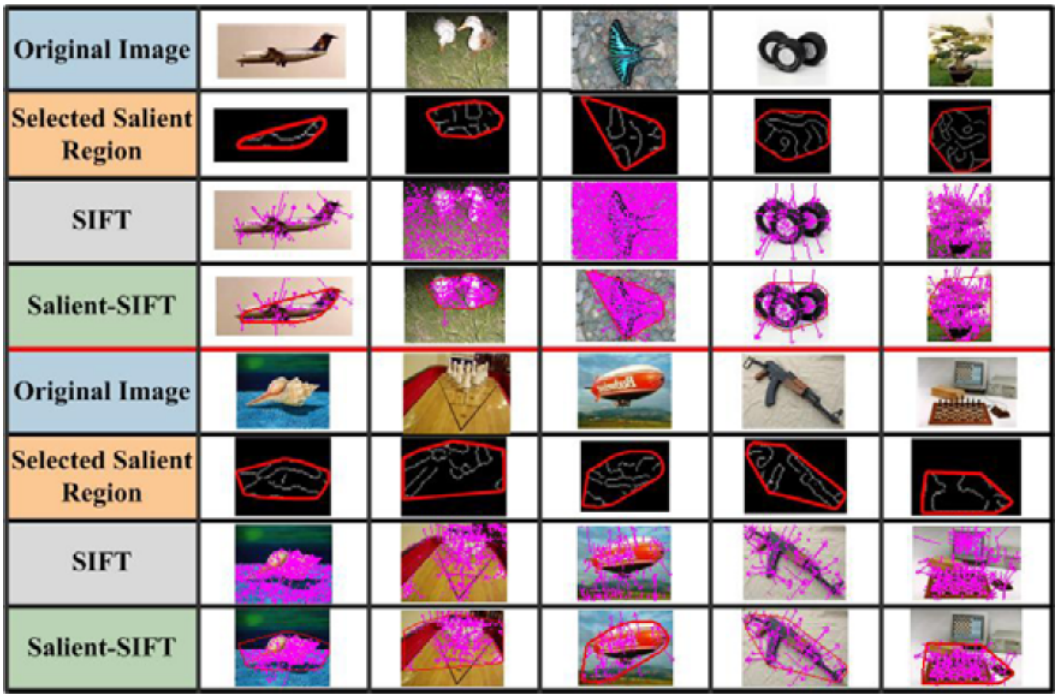

Figura 3.6: Exemplo dos descritores extraídos pelo *Salient*-SIFT. Ilustração retirada do artigo [Liang et al. 2010].

et al. 2007] e Oxford. Para a base de dados Caltech-256 [Griffin et al. 2007] a estratégia melhorou os resultados e na outra não houve melhora devido à disposição dos objetos de interesse nas imagens dessa base.

Figura 3.7: Exemplo dos pontos-chave selecionados de uma determinada imagem variando o *threshold* aplicado no mapa de saliência. Ilustração baseada no artigo [Nakamoto e Toriu 2011].

3.2 Considerações Finais

Com o intuito de melhorar cada vez mais o poder discriminativo dos vetores de características utilizados no processo de recuperação e/ou classificação de imagens, técnicas que se baseiam na percepção visual humana estão sendo utilizadas. Isso devido principalmente à facilidade de se definir qual a região de maior importância de uma determinada cena. Entretanto, generalizar a utilização dessas técnicas para as diversas aplicações é complicado pois por mais que se consiga separar regiões relevantes (*foreground*) e não relevantes (*background*) de uma determinada imagem, as relações entre essas regiões também podem ser importantes e influenciar nos resultados.

De acordo com os estudos realizados, os trabalhos encontrados na literatura consideram apenas o *foreground* como a parte mais importante da imagem a ser utilizada durante a recuperação e/ou classificação desconsiderando então as informações de *background*. Além disso, na maioria das técnicas, o *foreground* é separado do *background* de forma binária, o que pode gerar perda de informações durante o processo de recuperação e/ou classificação devido à incerteza em se classificar determinadas regiões da imagem como pertencentes ao *foreground* ou ao *background*.

Portanto, os métodos estudados motivaram a criação de técnicas que permitam ponderar a importância de características presentes no *foreground* e no *background* da ima-

gem durante o processo de recuperação, além de propor uma estratégia de separação do *background* e *foreground* evitando a perda de informações. Neste trabalho propomos a extensão da abordagem *bag-of-visual-features* alterando o processo de construção dos histogramas de frequência de palavras visuais. As técnicas propostas exploram a distinção entre o *background* e o *foreground* de duas maneiras diferentes: separação binária e separação *fuzzy*. Além disso, nessas duas abordagens, são explorados o uso de duas técnicas de extração de mapas de saliência, uma proposta por [Itti et al. 1998] e outra por [Harel et al. 2007]. Apesar da dificuldade em comparar e avaliar os métodos propostos devido à grande variedade de bases de dados utilizadas em trabalhos da literatura, foram realizados experimentos para a avaliar o desempenho das técnicas nas mais diversas variações de contexto levando em consideração a escala, rotação, cor, translação, iluminação e mudança de cena.

O próximo capítulo apresenta as técnicas propostas.

Capítulo 4

Proposta de descritores de características considerando a percepção visual humana

O *bag-of-visual-features* é uma estratégia de recuperação de imagens muito utilizada para o reconhecimento de objetos. É muito comum, na literatura, a utilização do descritor SIFT para representar esses objetos durante a fase de extração de características dessa estratégia. Isso se deve à sua eficiência na descrição das características de imagens, principalmente por ter a capacidade de ser invariante a mudanças de escala, rotação, orientação e iluminação.

Para qualquer objeto em uma imagem, pontos de interesse sobre o objeto são extraídos pelo SIFT para fornecer a descrição de características do mesmo. Estes pontos geralmente ficam em regiões de alto contraste da imagem, tais como as bordas do objeto. Uma vez geradas as características desse objeto, elas são utilizadas na abordagem *bag-of-visual-features* para realizar a discriminação do mesmo perante as outras imagens da base de dados.

Dependendo da imagem a ser analisada, o grande volume de características que o SIFT extrai de toda a imagem sem distinguir, por exemplo, o objeto de interesse (*foreground*) e o fundo (*background*) da imagem, pode prejudicar a representação desse objeto. Com isso, podemos perceber a importância do processo de distinção do *foreground* e *background* para melhorar a qualidade discriminativa de representação da imagem.

Estudos da percepção visual humana permitiram a criação de técnicas responsáveis por identificar partes relevantes e irrelevantes da imagem de forma automática. Essas técnicas são baseadas em extração de mapas de saliência que, quando aplicada a uma imagem, são capazes de identificar o que é relevante de acordo com a percepção humana. Algumas dessas técnicas são propostas por [Itti et al. 1998] e [Harel et al. 2007].

Este trabalho explora essa capacidade dos mapas de saliência em identificar, de forma

rápida e automática, as partes mais relevantes e menos relevantes de uma imagem com o intuito de melhorar o desempenho de recuperação de um determinado objeto ou cena. Para isso, considerando e estendendo a abordagem *bag-of-visual-features*, são propostos dois métodos para a montagem e processamento dos histogramas de palavras visuais, a construção de descritores usando o mapa de saliência em uma abordagem binária e a construção de descritores usando o mapa de saliência em uma abordagem *fuzzy*.

Neste capítulo é detalhado o funcionamento dessas duas propostas.

4.1 Uso do Mapa de Saliência para separar *foreground* e *background*

Os mapas de saliência (MS) podem ser vistos como uma função de pertinência nebulosa: os valores mais altos indicam pontos na imagem com maior atenção visual. Neste trabalho consideramos que os pontos de maior atenção serão referenciados com objetos ou regiões de interesse (*foreground*) visto que os pontos de menor atenção serão considerados como o fundo da imagem (*background*).

Podemos utilizar os mapas de saliência para separar *background* de *foreground* de duas maneiras diferentes: separação binária e separação nebulosa. Abaixo são descritos esses dois processos de separação:

- **Distinção binária entre *foreground* e *background*:** Nesta abordagem a função nebulosa que representa o mapa de saliência é binarizada, o *foreground* é representado por branco enquanto que o *background* é representado por preto.

  Para binarizar o mapa de saliência da imagem é criado um *threshold*. Considere $MS_{m,n}$ como sendo a matriz de valores reais do mapa de saliência. O *threshold* é gerado de acordo com uma média simples dos valores do mapa de saliência conforme mostra a Equação 4.1:

  $$threshold = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} MS(i,j)}{mn} \tag{4.1}$$

  A Figura 4.1 mostra a binarização dos mapas de saliência nos modelos de [Itti et al. 1998] e [Harel et al. 2007].

- **Distinção nebulosa (*fuzzy*) entre *foreground* e *background*:** A Figura 4.2 apresenta uma imagem de uma girafa e seu respectivo mapa de saliência com três pontos em destaque. Como as regiões mais claras do mapa de saliência representam o objeto de interesse (*foreground*) e a parte mais escura o fundo (*background*) da imagem podemos dizer que o ponto azul representa uma região de *foreground* e o

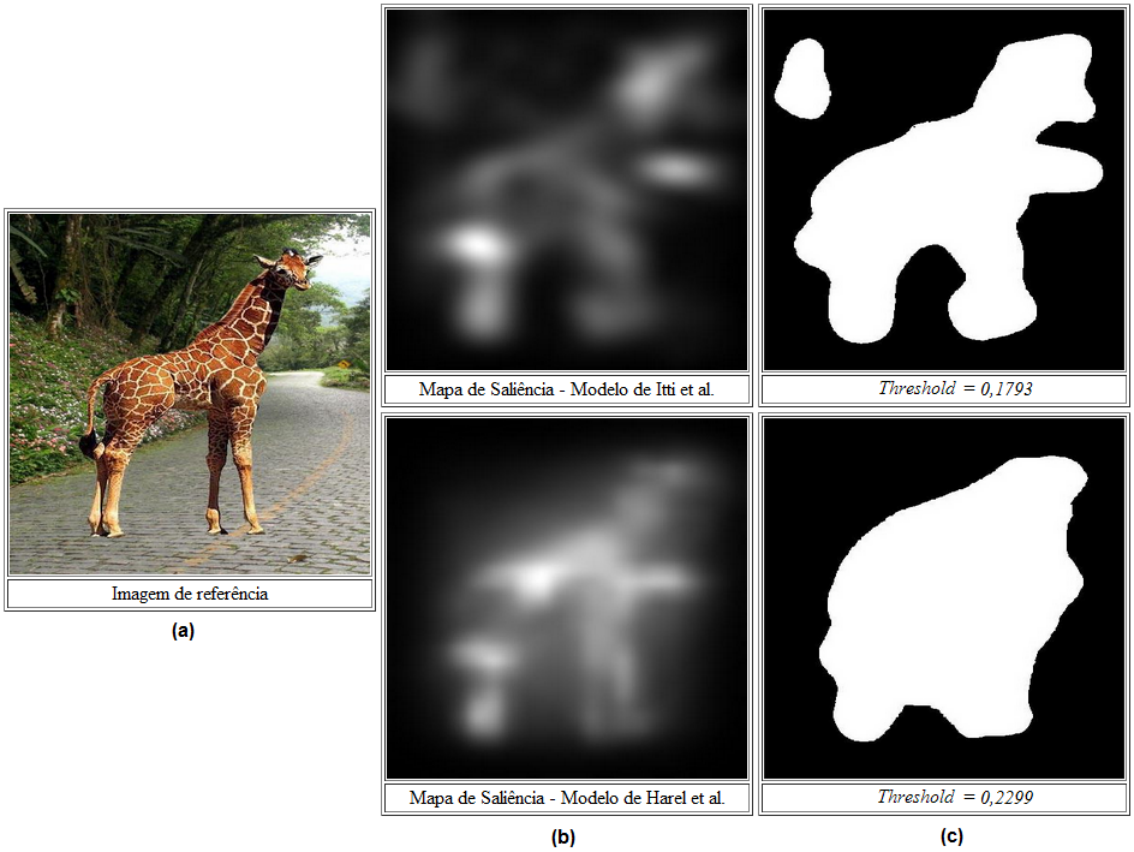

Figura 4.1: (a) imagem de referência; (b) mapas de saliência dos modelos de [Itti et al. 1998] e [Harel et al. 2007]; (c) imagens binárias dos dois modelos de extração de mapa de saliência, em que a parte branca representa o *foreground* e a parte preta representa o *background* da imagem.

ponto vermelho uma região de *background*. No entanto, temos incerteza em classificar o ponto amarelo, pois o mesmo está em uma região de transição entre a região de *foreground* e de *background* da imagem.

Com base nessa análise, propomos um método para classificar um *pixel* como *foreground* ou *background* capaz de modelar esse grau de incerteza utilizando a Teoria dos Conjuntos Nebulosos. Nessa teoria, um elemento pode pertencer a mais de um conjunto com diferentes graus de pertinência.

Assim, uma vez obtido o mapa de saliência (MS) de uma imagem, os valores do mapa são normalizados para o intervalo [0,1]. O mapa de saliência normalizado representa os graus de pertinência com que cada *pixel* da imagem pertence ao *foreground*, enquanto o seu complemento representa o grau de pertinência com que cada *pixel* pertence ao *background* da imagem.

O complemento de uma função de pertinência nebulosa é obtido de acordo com a Equação 4.2:

$$\overline{MS} = 1 - MS \tag{4.2}$$

Figura 4.2: Representação do mapa de saliência com três pontos em destaque em que o ponto azul representa uma região de *foreground*, o ponto vermelho uma região de *background* e o ponto amarelo uma região de transição entre o *background* e *foreground* da imagem.

Nas próximas seções são apresentadas duas maneiras diferentes de construir descritores de características considerando os processos de distinção binária e distinção *fuzzy* do *background* e *foreground* da imagem apresentados nessa seção.

4.2 Construção de descritores de características usando distinção binária entre *background* e *foreground*

Uma vez definido o processo binário de distinguir o *foreground* do *background* da imagem, conforme apresentado na seção 4.1, propomos dois descritores responsáveis pela extração de características utilizando os mapas de saliência, o *Binary Image Descriptor Based on Itti's Saliency Map* (BISMI) e o *Binary Image Descriptor Based on Harel's Saliency Map* (BISMH).

Ambos utilizam a abordagem *bag-of-visual-features*, com o descritor SIFT para representar os pontos-chave, o *k-means* para construir o dicionário de palavras visuais e o processo de distinção binária entre *foreground* e *background* da imagem utilizando os mapas de saliências para representar a localização espacial de palavras visuais na imagem. A diferença é que, para o processo de distinção binária, o BISMI utiliza o mapa de saliência proposto por [Itti et al. 1998] e o BISMH utiliza o mapa de saliência proposto por [Harel et al. 2007].

Considerando a abordagem *bag-of-visual-features*, uma vez extraídas as características

utilizando o SIFT e montado o dicionário de palavras visuais, são construídos os histogramas das imagens. BISMI e BISMH são responsáveis por montar estes histogramas selecionando as características presentes no *foreground* e no *background* da imagem.

Assim, para construir o histograma de frequência de palavras visuais, considerando o mapa de saliência na abordagem binária, inicialmente extraímos os mapas de saliência da imagem a ser tratada. O BISMI e o BISMH são compostos por dois histogramas: um que representa as palavras visuais que aparecem no *foreground* e outro que representa as palavras visuais que aparecem no *background* de uma imagem. Para toda palavra presente na imagem, é verificado se a mesma faz parte do *foreground* ou *background* da imagem. Caso faça parte do *foreground*, será incrementada a frequência dessa palavra no conjunto de palavras *foreground* do histograma, caso contrário, no conjunto de palavras *background* do histograma.

Essa distinção ocorre de acordo com o processo de limiarização utilizado para binarizar o mapa de saliência, conforme apresentado na seção 4.1. Toda palavra visual da imagem que possuir o valor do mapa de saliência menor que o *threshold* fará parte do conjunto de palavras do *background* da imagem e toda palavra da imagem que possuir o valor do mapa de saliência maior ou igual ao *threshold* fará parte do conjunto de palavras do *foreground* da imagem.

Os passos para se obter o BISMI e/ou o BISMH são descritos na sequência e eles encontram-se ilustrados na Figura 4.3. Seja I a imagem a ser analisada e D o dicionário de palavras visuais temos:

- Detectar todos os pontos-chave de I e derivar os descritores SIFT para cada um. Nesta etapa foi utilizada a detecção de pontos-chave e seu descritor SIFT correspondente como proposto por [Lowe 2004];
- Atribuir a cada descritor SIFT a palavra visual mais próxima em D;
- Calcular o mapa de saliência (MS) para I. Caso seja o descritor BISMI utiliza-se o método de extração de [Itti et al. 1998] e caso seja o BISMH utiliza-se o método de extração de [Harel et al. 2007];
- Calcular o *threshold* de binarização de MS conforme Equação 4.1;
- Montar o histograma H_f que representa a frequência de palavras visuais que aparecem no *foreground* de I. Toda palavra visual de I que possuir o valor de seu respectivo mapa de saliência maior ou igual ao *threshold* calculado fará parte de H_f;
- Montar o histograma H_b que representa a frequência de palavras visuais que aparecem no *background* de I. Toda palavra visual de I que possuir o valor de seu respectivo mapa de saliência menor que o *threshold* calculado fará parte de H_b;
- Montar o novo descritor de imagem BISMI ou BISMH concatenando H_f e H_b. A Figura 4.4 mostra a composição do histograma concatenado.

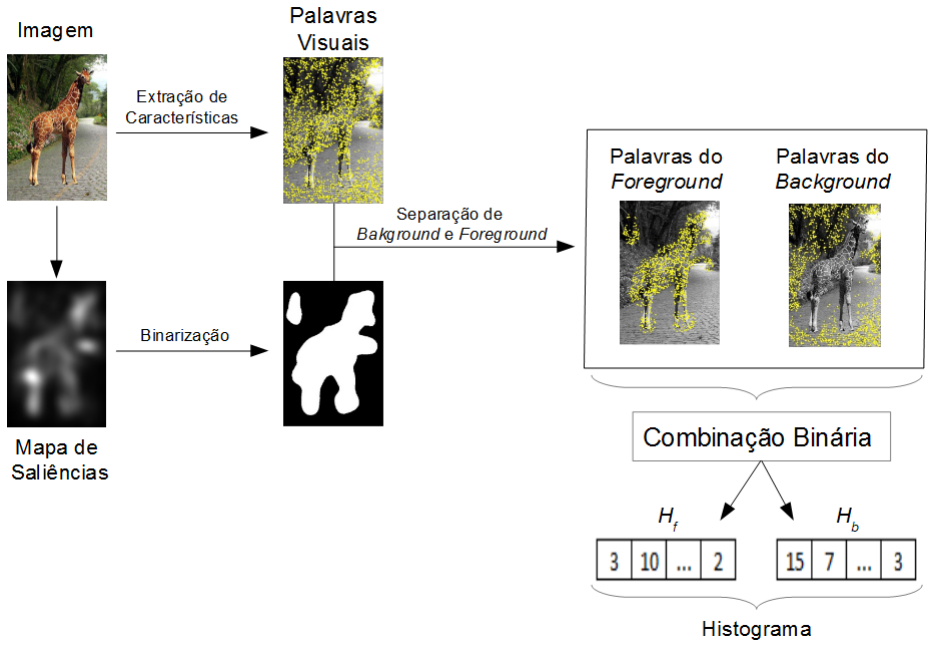

Figura 4.3: Esquema para a geração dos descritores BISMI e BISMH.

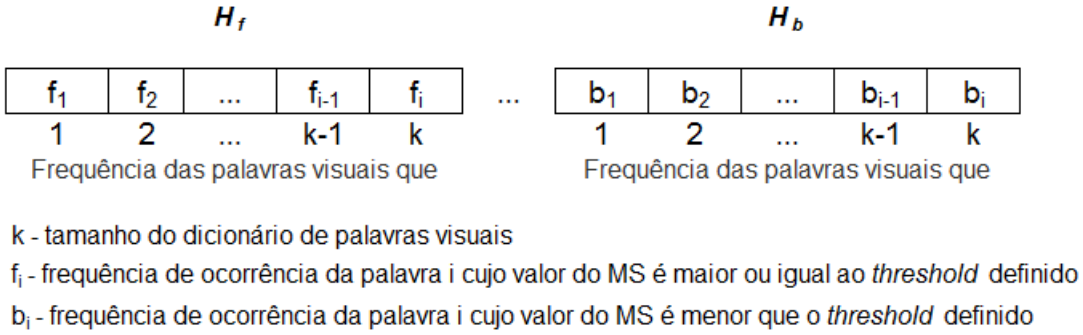

Figura 4.4: Composição e estrutura do histograma dos descritores BISMI e BISMH.

As Figuras 4.5 e 4.6 mostram exemplos da separação das palavras visuais em uma imagem considerando o modelo de [Itti et al. 1998] e o modelo [Harel et al. 2007] para montar os descritores de BISMI e BISMH respectivamente. Neste exemplo podemos perceber a importância em se realizar essa separação pois dessa forma conseguimos dar maior importância às palavras que mais discriminam o objeto em análise.

A separação dos conjuntos de frequência de palavras visuais de *background* e *foreground* permite que seja ponderada a importância dos dois conjuntos podendo, por exemplo, considerar o conjunto de *background* 70% mais importante que o conjunto de *foreground*. Neste caso o interesse da recuperação seria de imagens cujo o *background* fosse mais semelhante do que o *foreground* em relação à imagem de consulta.

Por fim, percebemos que um dos fatores limitantes dessa abordagem é a dificuldade em se encontrar um determinado *threshold* que otimize a melhor forma de se classificar a palavra em *foreground* ou *background*. Como foi mostrado anteriormente, para resolver esse problema [Nakamoto e Toriu 2011] propôs a variação do *threshold* adotando o que

tivesse melhor desempenho. Na próxima seção também propomos uma técnica capaz de superar essa dificuldade encontrada.

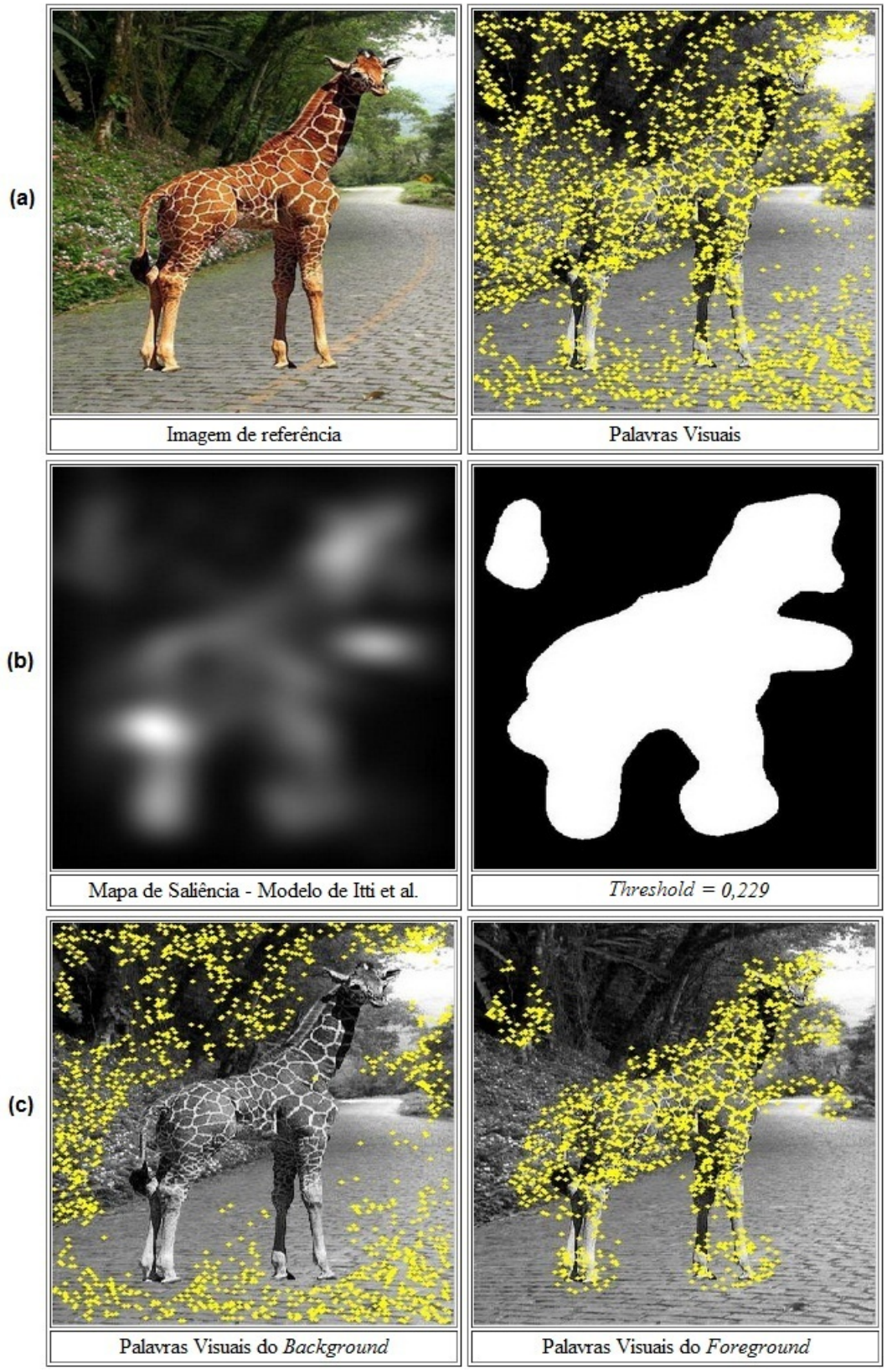

Figura 4.5: (a) imagem de análise e imagem de representação de todas as palavras visuais encontradas na mesma; (b) mapa de saliência extraído pelo modelo de [Itti et al. 1998] e imagem binária com aplicação do *threshold* no mapa; (c) palavras visuais que fazem parte do *background* e do *foreground* da imagem utilizadas para montar o histograma do descritor BISMI.

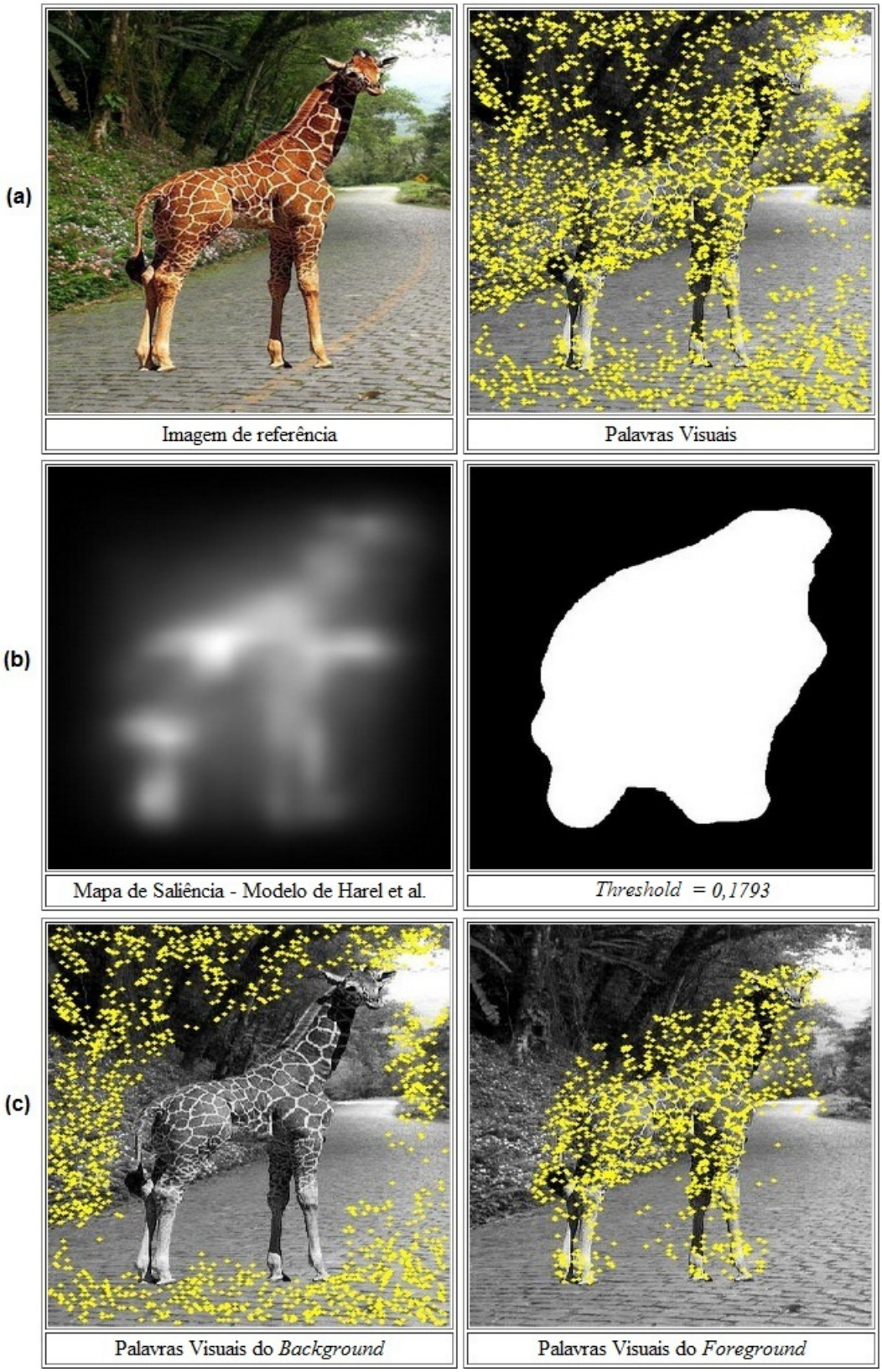

Figura 4.6: (a) imagem de análise e imagem de representação de todas as palavras visuais encontradas na mesma; (b) mapa de saliência extraído pelo modelo de [Harel et al. 2007] e imagem binária com aplicação do *threshold* no mapa; (c) palavras visuais que fazem parte do *background* e do *foreground* da imagem utilizadas para montar o histograma do descritor BISMH.

4.3 Construção de descritores de características usando distinção *fuzzy* entre *background* e *foreground*

Com o intuito de representar a localização espacial das palavras visuais apresentamos os descritores *Fuzzy Image Descriptor Based on Itti's Saliency Map* (FISMI) e o *Binary Image Descriptor Based on Harel's Saliency Map* (FISMH) com a proposta de extrair as características da imagem utilizando os mapas de saliência sem a necessidade de definir um *threshold* para separar o *background* do *foreground* da imagem. Ambos utilizam a abordagem *bag-of-visual-features* e o processo de distinção *fuzzy* entre *foreground* e *background* da imagem. A diferença é que, para o processo de distinção *fuzzy*, o FISMI utiliza o mapa de saliência proposto por [Itti et al. 1998] e o FISMH utiliza o mapa de saliência proposto por [Harel et al. 2007].

Assim como os descritores BISMI e BISMH, também é utilizado o descritor SIFT para representar os pontos-chave e o *k-means* para construir o dicionário de palavras visuais. O que difere é a maneira de representar o *foreground* e o *background* da imagem. FISMI e FISMH utilizam o processo de distinção *fuzzy* para essa representação gerando também dois histogramas, um que representa as palavras visuais que aparecem no *foreground* e outro que representa as palavras visuais que aparecem no *background* da imagem.

O mapa de saliência (MS) pode ser tratado como uma função de pertinência *fuzzy* em que os valores mais elevados indicam pontos da imagem com maior atenção visual, definida como $MS : I \rightarrow [0, 1]$. Conforme apresentado em seção anterior, utilizamos a função de pertinência *fuzzy* MS para representar o *foreground* e seu complemento para representar o *background*, dado por $\overline{MS} = 1 - MS$.

Os passos para se obter o FISMI e/ou o FISMH são descritos na sequência e eles encontram-se ilustrados na Figura 4.7. Seja I a imagem a ser analisada e D o dicionário de palavras visuais temos:

- Detectar todos os pontos-chave de I e derivar os descritores SIFT para cada um. Nesta etapa foi utilizada a detecção de pontos-chave e seu descritor SIFT correspondente como proposto por [Lowe 2004];
- Atribuir a cada descritor SIFT a palavra visual mais próxima em D;
- Calcular o mapa de saliência (MS) e seu complemento ($\overline{MS}$) para I. Caso seja o descritor FISMI utiliza-se o método de extração de [Itti et al. 1998] e caso seja o FISMH utiliza-se o método de extração de [Harel et al. 2007];
- Montar o histograma H_f que representa a frequência de palavras visuais que aparecem no *foreground* de I a partir da função de pertinência MS. Os elementos de H_f representam a frequência ponderada, em MS, de cada palavra visual que aparece no *foreground* de I.

- Montar o histograma H_b que representa a frequência de palavras visuais que aparecem no *background* de I a partir da função de pertinência $\overline{MS}$. Os elementos de H_b representam a frequência ponderada, em $\overline{MS}$, de cada palavra visual que aparece no *background* de I.
- Montar o novo descritor de imagem FISMI ou FISMH concatenando H_f e H_b. A Figura 4.4 mostra a composição do histograma concatenado.

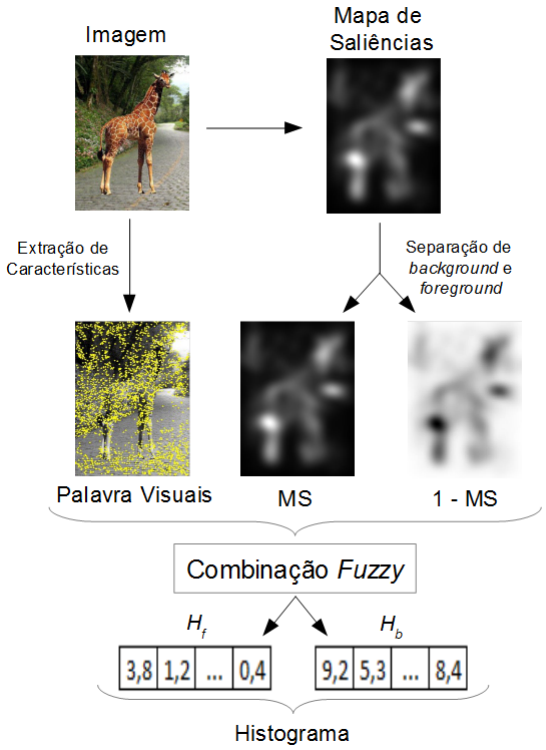

Figura 4.7: Esquema para a geração dos descritores FISMI e FISMH.

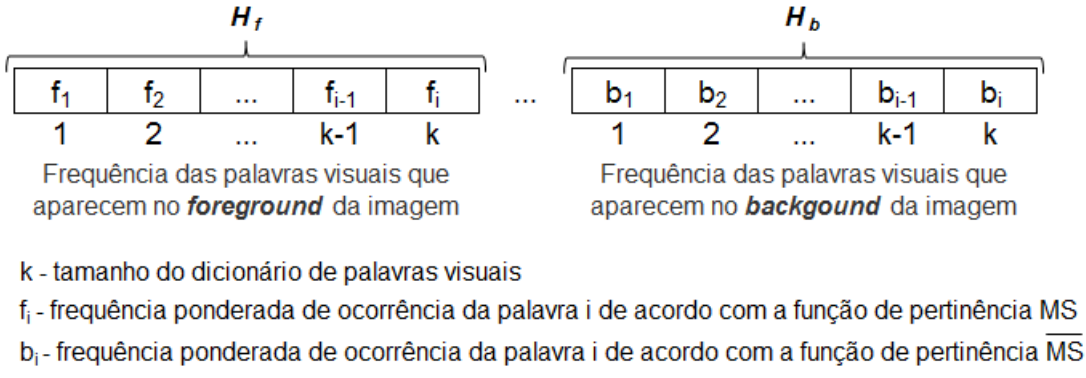

Figura 4.8: Composição e estrutura do histograma dos descritores FISMI e FISMH.

Como apresentado, FISMI e FISMH descrevem o *foreground* e o *background* das imagens separadamente. Dessa maneira, podemos atribuir pesos aos histogramas de representação para enfatizar a parte da consulta da imagem que é mais importante para a pesquisa de similaridade. Vale a pena notar que podemos procurar pelo *foregorund*, pelo *background* ou por uma combinação de ambos.

4.4 Considerações Finais

Neste capítulo apresentamos quatro descritores de características que se baseiam na percepção visual humana. A partir das técnicas de extração da atenção visual das imagens, representada pelos mapas de saliência, foi possível tratar de forma diferente o *background* e o *foreground* da mesma. Propusemos a distinção do *background* e do *foreground* de duas formas diferentes, a separação binária e a separação *fuzzy*. Na separação binária temos um fator limitante que é a escolha do *threshold* que otimize a melhor forma de separar o *background* do *foreground* da imagem. Em contrapartida, a separação *fuzzy* resolve este problema representando a incerteza presente nas regiões de transição entre o *background* e o *foreground*.

No próximo capítulo será descrito todos os experimentos realizados neste trabalho para avaliar os descritores propostos.

Capítulo 5

Experimentos

Neste Capítulo são apresentados experimentos que comparam os métodos propostos aos métodos baseados em cor e à abordagem *bag-of-visual-features* proposta por Csurka [Csurka et al. 2004]. Os experimentos foram realizados na base de dados Wang [Wang et al. 2001] e em bases de imagens *Near Duplicate*, as quais também chamamos de bases de versões, construídas especificamente para os testes.

5.1 Bases de dados

Com o intuito de avaliar as propostas deste trabalho, foram realizados vários testes. Para isso utilizaram-se bases de dados diferentes, a Wang [Wang et al. 2001] e 8 bases de versões. Nas próximas subseções serão apresentadas a base Wang e as bases de versões. Para as bases de versões, será descrita a metodologia de criação das mesmas por se tratarem de novas bases.

5.1.1 Base de dados Wang

A base de dados Wang [Wang et al. 2001] consiste em um subconjunto de 1000 imagens selecionadas manualmente do banco de dados de fotos da Corel. As imagens são categorizadas em 10 classes diferentes com 100 imagens cada uma. As 10 classes são usadas para estimativa de relevância: dada uma imagem de consulta, supõe-se que o usuário está a procura de imagens da mesma classe, e portanto, as restantes 99 imagens da mesma classe são consideradas relevantes e as imagens de todas as outras classes são consideradas irrelevantes. As classes dessa base de dados são: *Africa*, *Beach*, *Monuments*, *Buses*, *Food*, *Dinosaurs*, *Elephants*, *Flowers*, *Horses* e *Mountains*. A Figura 5.1 mostra uma imagem representante de cada classe da base Wang.

Figura 5.1: Exemplo de imagem de cada uma das 10 classes da base de dados Wang [Wang et al. 2001] mostrada juntamente com o nome de sua respectiva classe.

5.1.2 Bases de dados de versões

Uma possível aplicação deste trabalho é a utilização da percepção visual humana para identificação de versões de imagens. Entende-se por versão de uma imagem toda e qualquer nova imagem formada a partir de alterações da imagem original. Como exemplo de alterações podemos citar a mudança de cores, mudança de fundo, alteração na iluminação, rotação de objetos, entre outros. A recuperação de versões de uma determinada imagem origem é de interesse da polícia federal para identificar violação de *copyright*, ou adulteração de imagens. Assim, como não encontramos uma base de dados pública na literatura que possua essas características, foram criadas algumas bases que simulam possíveis alterações em imagens, gerando diversas versões.

Nessas novas bases foram utilizadas imagens de duas bases públicas muito utilizadas na literatura, a base Caltech 256 [Griffin et al. 2007] e a base Wang [Wang et al. 2001] (apresentada na subseção anterior). Ao todo foram criadas 8 bases com um total de 4720 imagens para que conseguíssemos avaliar o comportamento da nova proposta nos mais diversos contextos, gerando versões de imagens variando as cores, a iluminação, o fundo, a escala, a rotação e a translação.

Cada base contém 10 classes de imagens diferentes. As classes utilizadas são: *Africa*, *Ibis*, *Buses*, *Giraffe*, *Elephant*, *Motorbike*, *Horses*, *Dinosaurs*, *LightHouse* e *Teepe*. A Figura 5.2 mostra uma imagem de cada classe utilizada. Abaixo segue o detalhamento e a metodologia utilizada para a criação de cada base.

- **Escala:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando o tamanho do objeto de interesse na cena. A base possui 110 imagens, sendo 11 imagens por classe. Das 11 imagens, 1 é a imagem original, 5 são imagens nas quais o objeto de interesse diminui gradativamente em rela-

Figura 5.2: Imagens de consulta de cada classe da base de imagens.

ção ao objeto de interesse (*foreground*) da imagem original e 5 são imagens nas quais o objeto de interesse aumenta gradativamente em relação ao objeto de interesse (*foreground*) da imagem original. Os fatores de escala utilizados foram $e \in \{2, 4, 6, 8, 10, 1/2, 1/4, 1/6, 1/8, 1/10\}$. A Figura 5.3 mostra algumas imagens de uma classe dessa base.

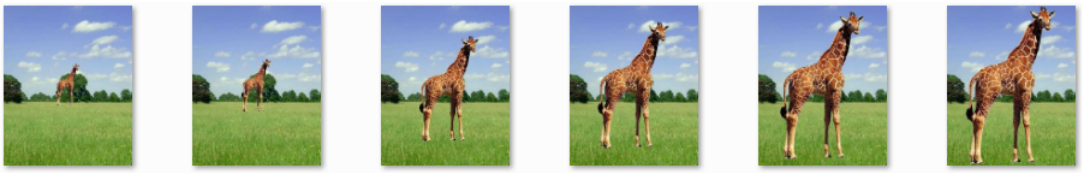

Figura 5.3: Imagens de exemplo da Base Escala.

- **Rotação:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando a rotação do objeto de interesse na cena. A base possui 390 imagens, sendo 39 imagens por classe. Das 39 imagens, 1 é a imagem original, 3 são as imagens rotacionadas por completo e 35 são as imagens onde o objeto de interesse é rotacionado na cena de 10° em 10° graus. A Figura 5.4 mostra algumas imagens de uma classe dessa base.

Figura 5.4: Imagens de exemplo da Base Rotação.

- **Translação:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando a posição do objeto de interesse na cena. A base possui 70 imagens, sendo 7 imagens por classe. Das 7 imagens, 1 é a imagem original e as outras 6

variam a posição do objeto de interesse em esquerda superior, esquerda inferior, direita superior, direita inferior, centro superior e centro inferior. A Figura 5.5 mostra as 7 imagens de uma classe dessa base.

Figura 5.5: Imagens de exemplo da Base Translação.

- **Cor:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando as cores da imagem. A base possui 550 imagens, sendo 55 imagens por classe. Das 55 imagens, 1 é a imagem original, 27 são geradas a partir da combinação dos valores 1.3, 1.5, 1.8 e as outras 27 também são geradas a partir da combinação dos valores 0.3, 0.5, 0.8. Os valores dessas combinações são multiplicados aos valores RGB de cada pixel da imagem. A Figura 5.6 mostra algumas imagens de uma classe dessa base.

Figura 5.6: Imagens de exemplo da Base Cor.

- **Iluminação:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando o contraste e o brilho das imagens. A base possui 370 imagens, sendo 37 imagens por classe. Das 37 imagens, 1 é a imagem original, 18 são as imagens que tiveram a variação de 0.1 a 0.9 para mais e para menos no brilho e 18 são as imagens que tiveram a variação de 0.1 a 0.9 para mais e para menos no contraste. A Figura 5.7 mostra algumas imagens de uma classe dessa base.
- **Fundo:** O objetivo dessa base é avaliar o comportamento das estratégias propostas variando o fundo das imagens. A base possui 110 imagens, sendo 11 imagens por classe. Das 11 imagens, 1 é a imagem original e 10 são imagens que tiveram o fundo da cena alterados. A Figura 5.8 mostra algumas imagens de uma classe dessa base.
- **Híbrido:** O objetivo dessa base é avaliar o comportamento das estratégias propostas juntando todas as variações realizada nas imagens das bases Escala, Rotação,

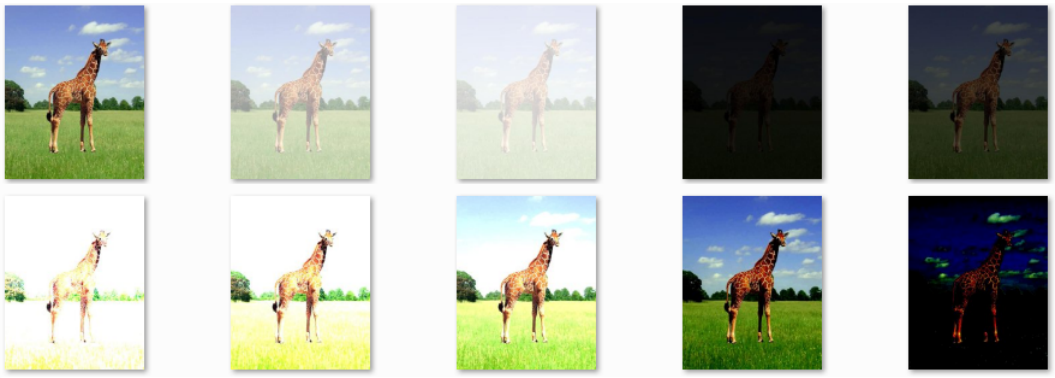

Figura 5.7: Imagens de exemplo da Base Iluminação.

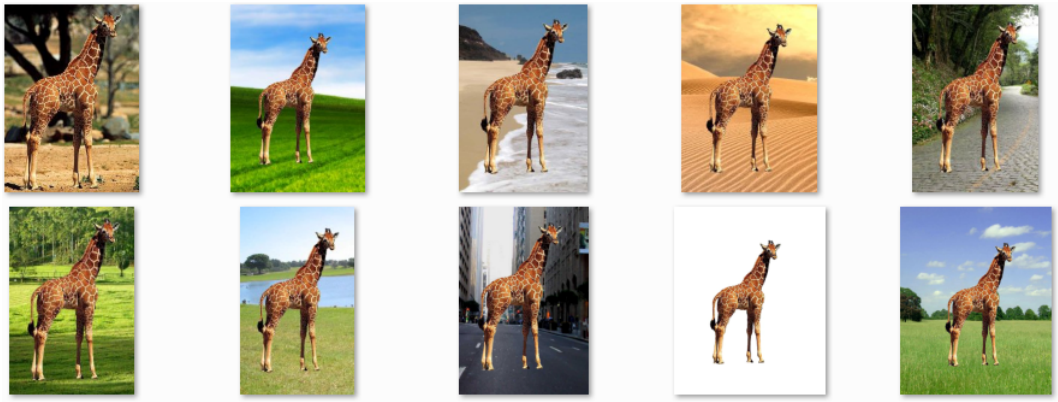

Figura 5.8: Imagens de exemplo da Base Fundo.

Translação, Cor, Iluminação e Fundo. A base possui 1550 imagens, sendo 155 imagens por classe. Das 155 imagens, 1 é a imagem original e as restantes são das outras bases. A Figura 5.9 mostra algumas imagens de uma classe dessa base.

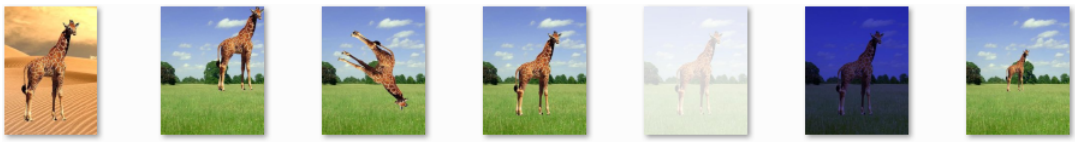

Figura 5.9: Imagens de exemplo da Base Híbrido.

- **Cena:** O objetivo dessa base é avaliar o comportamento das estratégias propostas quando se consulta pelo fundo da imagem. A base possui 1570 imagens, sendo 157 imagens por classe. Das 157 imagens, 1 é a imagem original, 1 é apenas a imagem da cena sem nenhum objeto, 1 é a imagem origem em que foi recortado o fundo a ser utilizado para montar as versões das imagens e as restantes são da base Híbrido. A Figura 5.10 mostra algumas imagens de uma classe dessa base.

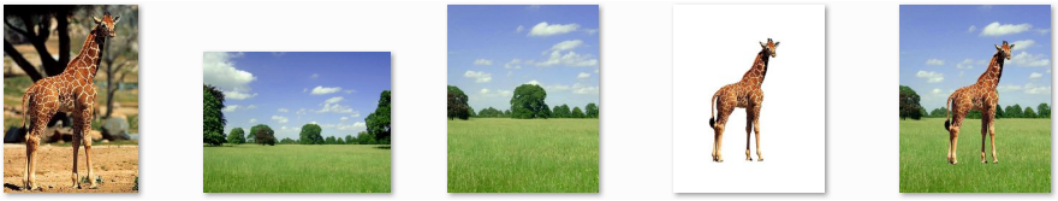

Figura 5.10: Imagens de exemplo da Base Cena.

5.2 Preparação dos Experimentos

Foram realizados os experimentos utilizando a base de dados Wang e as bases de dados de versões, descritas na seção 5.1. Utilizando essas bases, conseguimos avaliar os métodos propostos quanto a capacidade de invariância às transformações geométricas, às mudanças de iluminação, às mudanças de cor, entre outros.

Independente da base de dados utilizada, foi implementado o processo de *bag-of-visual-features* conforme mostrado no Capítulo 2, e para cada fase dessa abordagem foram utilizadas as seguintes configurações:

1. **Extração de Características:** Nesta fase utilizou-se o SIFT [Lowe 2004] como descritor dos pontos de interesse das imagens.
2. **Construção do Dicionário de Palavras Visuais:** Para construir o dicionário visual utilizou-se o método de agrupamento *k-means* [Jain et al. 1999]. Foram realizados testes variando o tamanho do dicionário (k) utilizando-se o tamanho que teve melhores resultados, o de 4000 palavras visuais. Devido ao grande volume de características, foram selecionadas, de forma aleatória, até 6000 características por classe de imagens da base utilizada para realizar o agrupamento. É importante ressaltar também que não foram utilizadas as características das imagens de consulta para formar o dicionário.
3. **Construção dos Histogramas de Palavras Visuais:** Nesta fase variou-se a maneira de se construir os histogramas de palavras visuais. Foram três formas utilizadas, a proposta por Csurka [Csurka et al. 2004] montando o histograma de acordo com a frequência que uma determinada palavra visual aparece na imagem e as abordagens que utilizam a percepção visual humana para montar os histogramas considerando a forma binária e *fuzzy*. Em todos os três casos utilizou-se a distância Euclidiana para identificar a que palavra visual determinado descritor de característica é mais similar.
4. **Busca por Similaridade:** O operador de similaridade utilizado para montar a lista de imagens similares a uma determinada consulta foi o KNNQ onde a quantidade de elementos recuperados é igual a quantidade de imagens que a classe da imagem de consulta possui. A distância de similaridade utilizada nesse operador foi a distância Cosseno.

Abaixo são apresentados os métodos analisados e comparados nos experimentos:

- **BaselineBoVF:** Método para recuperação de imagem baseada em conteúdo proposta por Csurka [Csurka et al. 2004] representando o baseline de comparação com o método *bag-of-visual-features*. Utilizou-se o descritor SIFT proposto por [Lowe 2004] e a imagem foi representada pelo histograma de frequência de palavras visuais;
- **SCD:** Descritor de características de cor proposto por [Manjunath et al. 2001] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **CLD:** Descritor de características de cor proposto por [Manjunath et al. 2001] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **EHD:** Descritor de características de cor proposto por [Manjunath et al. 2001] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **CEDD:** Descritor de características composto (cor e textura) proposto por [Chatzichristofis et al. 2010] e descrito no Capítulo 2. Os resultados foram obtidos utilizando a ferramenta chamada LIRE [Lux e Chatzichristofis 2008] e adequando-os a metodologia de experimentação utilizada.
- **FCTH:** Também descritor de características composto (cor e textura) proposto por [Chatzichristofis et al. 2010] e descrito no Capítulo 2. Os resultados foram obtidos utilizando a ferramenta chamada LIRE [Lux e Chatzichristofis 2008] e adequando-os a metodologia de experimentação utilizada;
- **BIC:** Descritor de características de cor proposto por [Stehling et al. 2002] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **LCPC:** Descritor de características de cor proposto por [Kimura et al. 2011] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **LCPC+EDGE:** Descritor de características composto (cor + forma) proposto por [Kimura et al. 2011] e descrito no Capítulo 2. Os resultados foram obtidos pelo trabalho de [Kimura et al. 2011] utilizando a mesma metodologia da experimentação realizada;
- **BISMI:** Construção do histograma de palavras visuais considerando os descritores que usam o mapa de saliência na abordagem binária descrito no Capítulo 4. É utilizada a técnica de [Itti et al. 1998] para extrair os mapas de saliência das imagens;

- **BISMH:** Construção do histograma de palavras visuais considerando os descritores que usam o mapa de saliência na abordagem binária descrito no Capítulo 4. É utilizada a técnica de [Harel et al. 2007] para extrair os mapas de saliência das imagens;
- **FISMI:** Construção do histograma de palavras visuais considerando os descritores que usam o mapa de saliência na abordagem *fuzzy* descrito no Capítulo 4. É utilizada a técnica de [Itti et al. 1998] para extrair os mapas de saliência das imagens;
- **FISMH:** Construção do histograma de palavras visuais considerando os descritores que usam o mapa de saliência na abordagem *fuzzy* descrito no Capítulo 4. É utilizada a técnica de [Harel et al. 2007] para extrair os mapas de saliência das imagens.

5.2.1 Configurações específicas para a base de dados Wang

Para a base de dados Wang, os métodos avaliados foram: SCD, CLD, EHD, CEDD, FCTH, BIC, LCPC, LCPC+EDGE, BaselineBoVF, BISMI, BISMH, FISMI e FISMH. Nos experimentos uma imagem de cada classe é utilizada como consulta, totalizando 10 consultas por método avaliado. A Figura 5.11 mostra as imagens de cada classe utilizadas nas consultas.

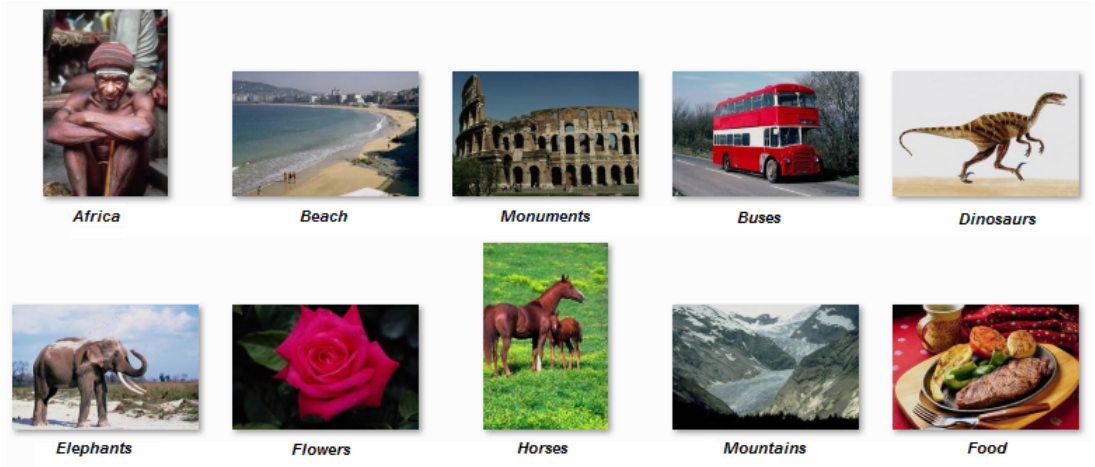

Figura 5.11: Imagem de cada classe utilizada nas consultas dos experimentos para a base de dados Wang.

5.2.2 Configurações específicas para as bases de versões

Para a base de dados de versões, os métodos avaliados foram: BaselineBoVF, BISMI, BISMH, FISMI, FISMH, CEDD e FCTH.

Todas esses métodos foram avaliados nas bases Cor, Iluminação, Escala, Rotação, Translação, Fundo, Híbrido e Cena conforme apresentadas e detalhadas na subseção 5.1.2. Em cada base variou-se a quantidade de classes utilizadas no experimento de duas em duas,

ou seja, cada base possui 10 classes (*Africa*, *Ibis*, *Buses*, *Giraffe*, *Elephant*, *Motorbike*, *Horses*, *Dinosaurs*, *LightHouse* e *Teepe*), assim, foram realizados experimentos com 2 (*Africa* e *Ibis*), 4 (*Africa*, *Ibis*, *Buses* e *Giraffe*), 6 (*Africa*, *Ibis*, *Buses*, *Giraffe*, *Elephant* e *Motorbike*), 8 (*Africa*, *Ibis*, *Buses*, *Giraffe*, *Elephant*, *Motorbike*, *Horses* e *Dinosaurs*) e 10 (*Africa*, *Ibis*, *Buses*, *Giraffe*, *Elephant*, *Motorbike*, *Horses*, *Dinosaurs*, *LightHouse* e *Teepe*) classes para todas as abordagens e em todas as bases.

Nos experimentos, uma imagem de cada classe é utilizada como consulta. A Figura 5.12 mostra as imagens de consulta de cada classe utilizadas nas bases Cor, Iluminação, Escala, Rotação, Translação e Híbrido.

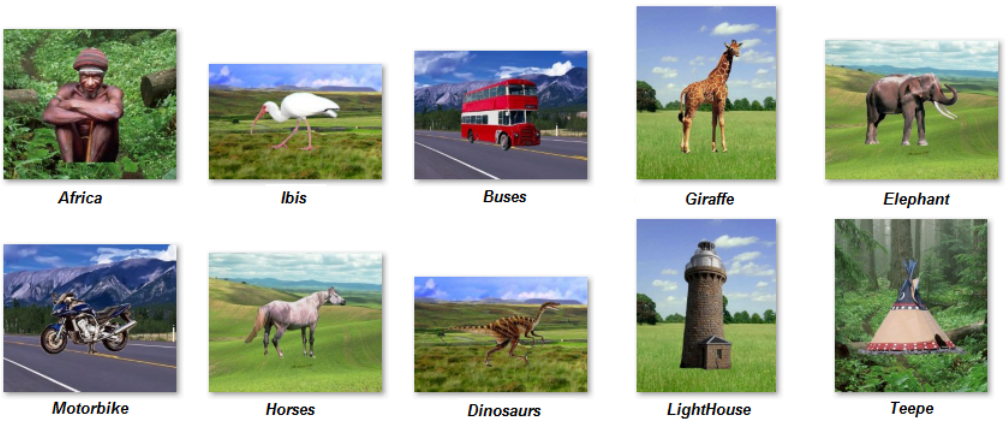

Figura 5.12: Imagens utilizadas como consulta de cada classe para as bases Cor, Iluminação, Escala, Rotação, Translação e Híbrido.

As imagens de consulta da base Fundo são mostradas pela Figura 5.13. Para cada classe utilizaram-se três imagens de consulta, a imagem originada das bases públicas, a imagem contendo somente o objeto de interesse e a imagem gerada com um fundo diferente da imagem original.

Para a base Cena, as imagens utilizadas como consultas são mostradas pela Figura 5.14.

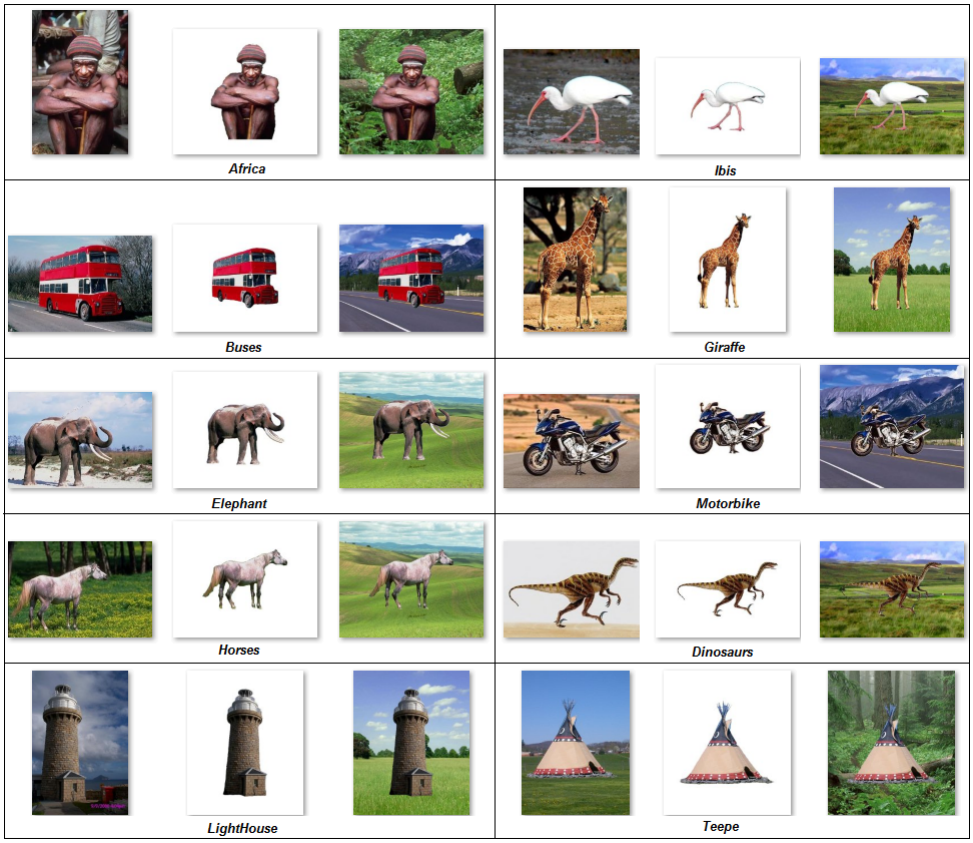

Figura 5.13: Imagens utilizadas como consulta de cada classe para a base Fundo.

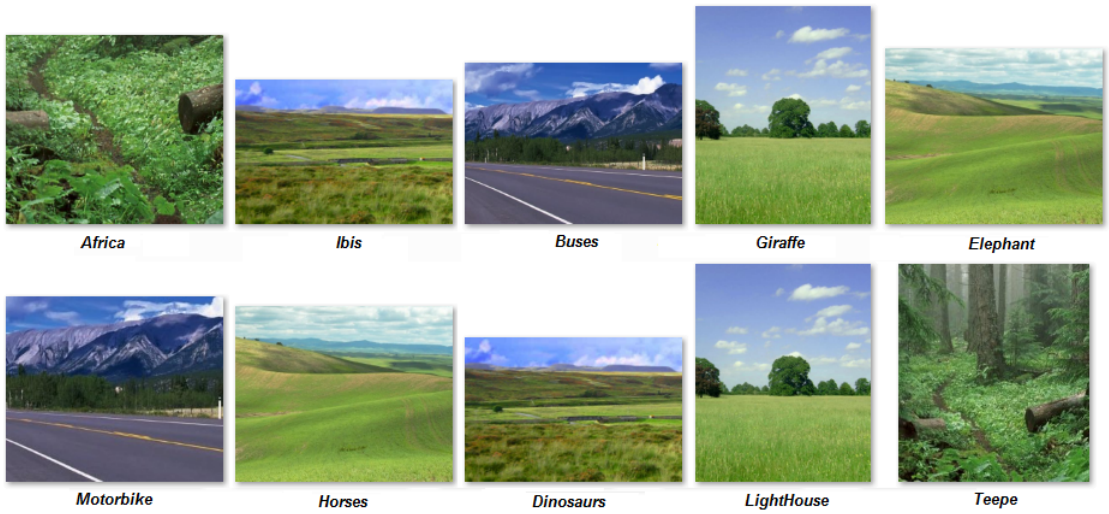

Figura 5.14: Imagens utilizadas como consulta de cada classe para a base Cena.

5.3 Avaliação dos Resultados

As medidas de avaliação utilizadas em cada experimento foram o MAP, para 100% de revocação e o $NDCG_k$ com o k sendo igual ao tamanho das classes da base em análise. Para uma determinada base, após calculado o MAP e o $NDCG_k$ para cada classe, é feita uma média aritmética das medidas para gerar o MAP e o $NDCG_k$ da abordagem em análise aplicada a base. Para cada base de imagens utilizada (8 no total), com exceção

da base de dados Wang, foram gerados gráficos mostrando a variação da medida $NDCG_k$ e MAP em relação ao aumento do número de classes utilizadas.

As próximas subseções apresentam os resultados obtidos para cada base de dados utilizada.

5.3.1 Base de dados Wang

Nesta subseção são apresentados os resultados dos experimentos realizados na base de dados Wang. Os resultados apresentados para os métodos SCD, CLD, EHD, CEDD, FCTH, BIC, LCPC, LCPC+EDGE foram obtidos através do trabalho de [Kimura et al. 2011]. Já os resultados dos métodos BaselineBoVF, BISMI, BISMH, FISMI e FISMH foram obtidos pela execução dos experimentos na base de dados Wang considerando a mesma metodologia utilizada por Kimura et al., inclusive as mesmas imagens de consulta, obtidas junto ao autor. Assim, a Tabela 5.1 mostra os valores de MAP em ordem decrescente para os métodos avaliados na base de dados Wang.

Tabela 5.1: Valores de MAP em ordem decrescente para a base de dados Wang.

Métodos	MAP
LCPC+Edge	0.67
LCPC	0.62
BIC	0.59
CEDD	0.59
FCTH	0.58
CLD	0.51
BISMI	0.42743
FISMI	0.42694
BISMH	0.42373
FISMH	0.41012
BaselineBoVF	0.39299
EHD	0.39
SCD	0.38

Conforme apresentado, os métodos LCPC+Edge e LCPC se destacaram na avaliação dessa base de dados seguidos dos métodos BIC, CEDD e FCTH. Para a base Wang percebemos que os métodos propostos (BISMI, BISMH, FISMI e FISMH) não foram tão efetivos apesar de terem obtido resultados melhores que BaselineBoVF, EHD e SCD. A proposta dos métodos desenvolvidos nesse trabalho utilizam como extrator de características o SIFT. A base de dados Wang não é muito bem representada pela extração dos descritores SIFT comparada a extração de características de cor. Outro ponto importante é que a abordagem *bag-of-visual-features* se comporta melhor na recuperação de objetos bem definidos, o que não acontece para determinadas classes da base de dados Wang, como por exemplo, a classe *Africa*, *Beach*, *Buildings*, *Mountains* e *Food*. A diversidade de

características das imagens dessas classes é muito grande. Na próxima seção os métodos propostos são avaliados em 8 bases de versões.

5.3.2 Bases de dados de versões

Nesta subseção são apresentados os resultados dos experimentos realizados nas 8 bases de versões criadas. Para cada base foram executados e analisados os experimentos para os métodos BaselineBoVF, BISMI, BISMH, FISMI, FISMH, CEDD e FCTH. Seguem os resultados obtidos:

- **Base de Cor:** No primeiro experimento foi avaliada o desempenho dos métodos na base Cor, em que as versões das imagens tiveram as suas cores alteradas em relação a imagem original. A Figura 5.15 apresenta a tabela de valores e o gráfico das medidas $NDCG_k$ e MAP. Podemos perceber que os quatro métodos propostos foram superiores às abordagens BaselineBoVF, CEDD e FCTH sendo que a abordagem FISMH foi a superior obtendo um ganho de até 11% em relação ao BaselineBoVF considerando a medida MAP. O baixo desempenho dos descritores de cor CEDD e FCTH é explicada devido à similaridade de cores entre as imagens de classes diferentes.

COR - $NDCG_{55}$					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,95175	0,94899	0,96199	0,90596	0,84698
BISMI	0,98812	0,98809	0,99009	0,95874	0,94437
BISMH	0,99407	0,98811	0,99208	0,95541	0,93138
FISMI	1,00000	0,99108	0,99207	0,97118	0,93380
FISMH	1,00000	0,99406	0,99208	0,97884	0,95036
CEDD	0,75034	0,67814	0,56097	0,46209	0,43572
FCTH	0,78253	0,69068	0,56856	0,46991	0,43118

COR - MAP					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,96932	0,96784	0,97807	0,92569	0,86777
BISMI	0,99921	0,99408	0,99422	0,97822	0,96233
BISMH	0,99984	0,99657	0,99722	0,97363	0,95467
FISMI	1,00000	0,99800	0,99741	0,98392	0,95421
FISMH	1,00000	0,99871	0,99842	0,98816	0,96705
CEDD	0,80287	0,65518	0,49660	0,38722	0,33557
FCTH	0,84057	0,67237	0,51638	0,40072	0,34951

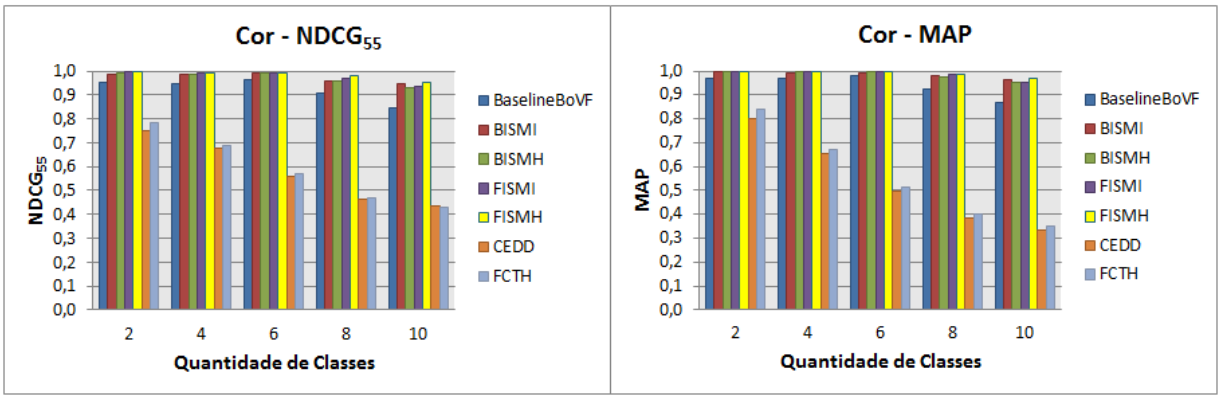

Figura 5.15: Resultados para a base de Cor variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 55$ considerando o total de imagens relevantes deste experimento.

- **Base de Iluminação:** A Figura 5.16 mostra os gráficos de $NDCG_k$ e MAP para a base Iluminação. Inicialmente, para duas classes, as abordagens CEDD e FCTH se

compara com as demais abordagens, no entanto, ao aumentar o número de classes utilizadas, esse desempenho cai bastante. É possível perceber a superioridade das abordagens que utilizam a percepção visual, mesmo aumentando a quantidade de classes utilizadas. A técnica FISMH foi a superior obtendo um ganho de até 17% em relação ao BaselineBoVF considerando a medida *MAP*.

ILUMINAÇÃO - $NDCG_{37}$					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,80231	0,77350	0,77814	0,75771	0,71699
BISMI	0,85051	0,84540	0,85001	0,81510	0,79129
BISMH	0,85940	0,85920	0,85023	0,81496	0,78819
FISMI	0,84160	0,82761	0,84421	0,81621	0,78489
FISMH	0,84193	0,85494	0,85348	0,82591	0,79808
CEDD	0,82975	0,70527	0,62252	0,54518	0,53690
FCTH	0,84714	0,69602	0,63217	0,54489	0,52880

ILUMINAÇÃO - MAP					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,85721	0,77707	0,77322	0,71635	0,64677
BISMI	0,89421	0,84879	0,83667	0,79252	0,74303
BISMH	0,90119	0,85744	0,84105	0,79373	0,74678
FISMI	0,89471	0,84072	0,83728	0,79551	0,74323
FISMH	0,89458	0,85302	0,84338	0,80547	0,76126
CEDD	0,87440	0,69405	0,56100	0,44695	0,40448
FCTH	0,88576	0,68445	0,57303	0,45516	0,40196

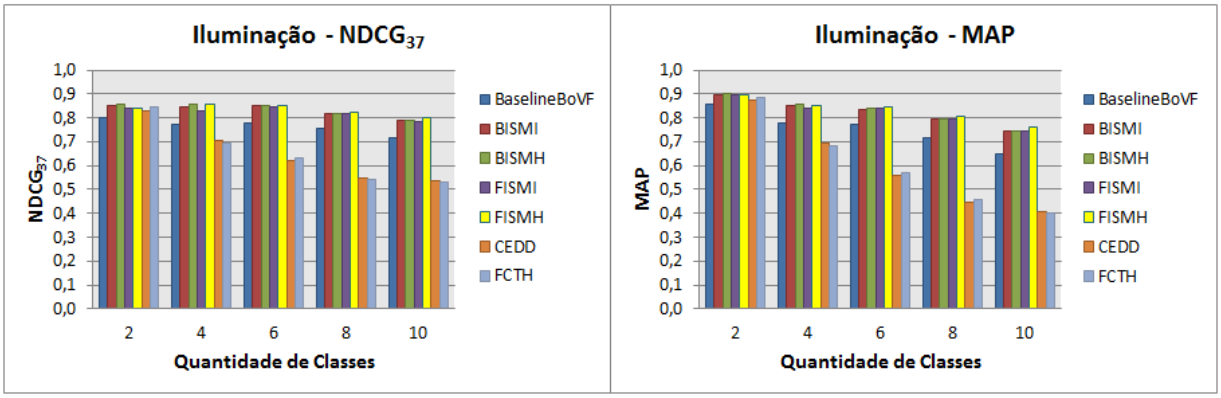

Figura 5.16: Resultados para a base de Iluminação variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 37$ considerando o total de imagens relevantes deste experimento.

- **Base de Escala, Rotação e Translação:** Com o objetivo de avaliar a invariância das abordagens em relação variações geométricas das imagens, foram realizados experimentos nas bases de Escala, Rotação e Translação cujo resultados são mostrados pelas Figuras 5.17, 5.18 e 5.19, respectivamente. Como era de se esperar, o desempenho das abordagens para estes experimentos tiveram pouca queda ou até mesmo nenhuma ao se aumentar o número de classes utilizadas. As abordagens CEDD e FCTH foram as que mais tiveram queda de desempenho, seguidos da abordagem BaselineBoVF. Em contrapartida, as abordagens que utilizam a percepção visual tiveram pouca perda, ou até mesmo nenhuma, como foi o caso das abordagens FISMI e FISMH para os experimentos nas bases de Rotação e Translação. Destaque mesmo para a abordagem FISMH que nos três experimentos se manteve quase que constante.

- **Base de Fundo:** A Figura 5.20 apresenta os resultados para a base Fundo. Neste experimento conseguimos avaliar a importância da utilização da técnica de extração da percepção visual para realizar a recuperação de imagens. Os métodos propostos

ESCALA - $NDCG_{11}$						ESCALA - MAP					
Método	Quantidade de Classes					Método	Quantidade de Classes				
	2	4	6	8	10		2	4	6	8	10
BaselineBoVF	1,00000	1,00000	0,99131	0,91516	0,88457	BaselineBoVF	1,00000	1,00000	0,99767	0,94711	0,91976
BISMI	1,00000	1,00000	1,00000	0,95224	0,95632	BISMI	1,00000	1,00000	1,00000	0,97838	0,97904
BISMH	1,00000	1,00000	1,00000	0,96482	0,96084	BISMH	1,00000	1,00000	1,00000	0,97828	0,97767
FISMI	1,00000	1,00000	1,00000	0,97206	0,95658	FISMI	1,00000	1,00000	1,00000	0,98623	0,98142
FISMH	1,00000	1,00000	1,00000	0,98485	0,97823	FISMH	1,00000	1,00000	1,00000	0,99108	0,99358
CEDD	1,00000	1,00000	0,93387	0,81441	0,81604	CEDD	1,00000	1,00000	0,95154	0,85355	0,85760
FCTH	1,00000	1,00000	0,95406	0,83386	0,79965	FCTH	1,00000	1,00000	0,97329	0,87340	0,84715

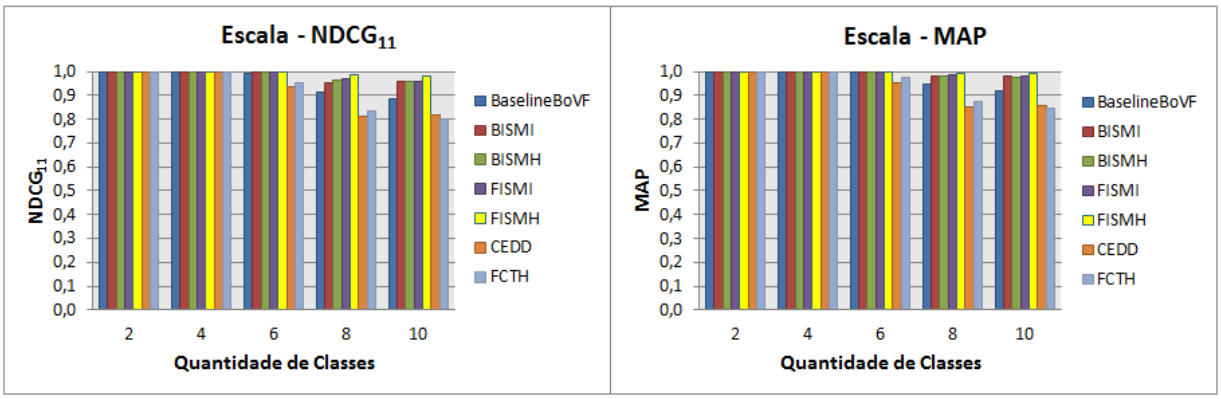

Figura 5.17: Resultados para a base de Escala variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 11$ considerando o total de imagens relevantes deste experimento.

que utilizaram essas técnicas tiveram superioridade sobre as abordagens BaselineBoVF, CEDD e FCTH. Isso se deve a capacidade em se separar o *background* do *foreground* da imagem. Nessa base, o interesse era de recuperar imagens que possuíssem os objetos de interesse (*foreground*) mais similares desconsiderando então a mudança de fundo (*background*) na imagem. As abordagens CEDD e FCTH, que utilizam as características de cor para calcular a similaridade das imagens, tiveram enorme dificuldade na recuperação devido à variedade de cores inseridas nas imagens ocasionadas pela variação dos fundos. Já a abordagem FISMH obteve um ganho de até 24% comparado a abordagem BaselineBoVF.

- **Base Híbrida:** O experimento utilizando a base Híbrida possibilitou reunir todas as mudanças de versões das imagens em apenas uma base para que fosse avaliada o desempenho das abordagens em coleções de imagens com alto grau de variação. Mesmo com essa alta diversificação das alterações das imagens, as abordagens BISMI, FISMI, BISMH e FISMH responderam melhor do que as abordagens BaselineBoVF, CEDD e FCTH conforme mostrado na Figura 5.21. A abordagem FISMH chegou a ter um ganho de até 15% em relação a abordagem BaselineBoVF.
- **Base de Cena:** Por fim, o último experimento realizado utilizou a base Cena com o intuito de avaliar a flexibilidade dos métodos propostos neste trabalho em permitir ao usuário recuperar imagens cuja cena é mais importante do que o objeto da imagem. Dessa forma a ponderação maior para o *background* do que para o *foreground*

ROTAÇÃO - $NDCG_{39}$						ROTAÇÃO - MAP					
Método	Quantidade de Classes					Método	Quantidade de Classes				
	2	4	6	8	10		2	4	6	8	10
BaselineBoVF	1,00000	1,00000	1,00000	0,99589	0,99839	BaselineBoVF	1,00000	1,00000	1,00000	0,99928	0,99994
BISMI	1,00000	1,00000	1,00000	0,99382	0,99666	BISMI	1,00000	1,00000	1,00000	0,99859	0,99919
BISMH	1,00000	1,00000	1,00000	0,98026	0,98986	BISMH	1,00000	1,00000	1,00000	0,98726	0,99473
FISMI	1,00000	1,00000	1,00000	0,99790	1,00000	FISMI	1,00000	1,00000	1,00000	0,99934	1,00000
FISMH	1,00000	1,00000	1,00000	1,00000	1,00000	FISMH	1,00000	1,00000	1,00000	1,00000	1,00000
CEDD	1,00000	1,00000	0,96995	0,93839	0,92180	CEDD	1,00000	1,00000	0,97888	0,94195	0,93492
FCTH	1,00000	1,00000	1,00000	0,96640	0,96482	FCTH	1,00000	1,00000	1,00000	0,97341	0,97511

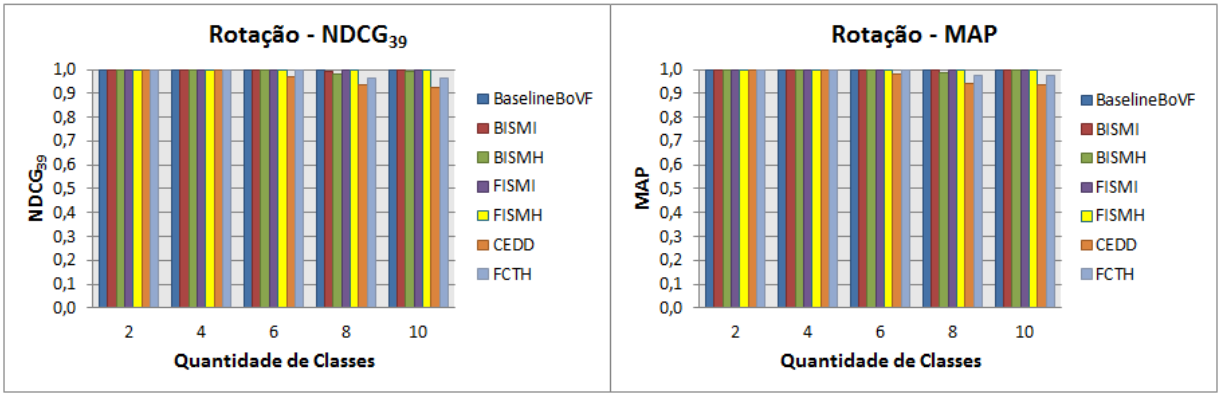

Figura 5.18: Resultados para a base de Rotação variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 39$ considerando o total de imagens relevantes deste experimento.

permite realizar consultas com esse objetivo. Assim, como mostra a Figura 5.22, é possível perceber que as abordagens propostas BISMI, FISMI, BISMH e FISMH além de permitir a flexibilidade de consulta ao usuário obtiveram melhores resultados do que a abordagem BaselineBoVF, CEDD e FCTH. Para este experimento não foi montado o gráfico de $NDCG_k$ porque o *ground-truth* das classes é variável conforme a quantidade de classes utilizadas no experimento. Como mostrado na Figura 5.14 algumas classes possuem a mesma imagem de cena aumentando então o *ground-truth* de ambas as classes ao serem utilizadas em conjunto, com isso, não foi possível calcular a média dos valores de $NDCG_k$ obtidos.

TRANSLAÇÃO - $NDCG_7$						TRANSLAÇÃO - MAP					
Método	Quantidade de Classes					Método	Quantidade de Classes				
	2	4	6	8	10		2	4	6	8	10
BaselineBoVF	1,00000	1,00000	1,00000	0,95557	0,95445	BaselineBoVF	1,00000	1,00000	1,00000	0,97099	0,97064
BISMI	1,00000	1,00000	1,00000	0,98749	0,98011	BISMI	1,00000	1,00000	1,00000	0,99224	0,98415
BISMH	1,00000	1,00000	1,00000	0,97842	0,99173	BISMH	1,00000	1,00000	1,00000	0,98542	0,99683
FISMI	1,00000	1,00000	1,00000	1,00000	1,00000	FISMI	1,00000	1,00000	1,00000	1,00000	1,00000
FISMH	1,00000	1,00000	1,00000	1,00000	1,00000	FISMH	1,00000	1,00000	1,00000	1,00000	1,00000
CEDD	1,00000	1,00000	1,00000	0,90533	0,90426	CEDD	1,00000	1,00000	1,00000	0,94182	0,93170
FCTH	1,00000	1,00000	1,00000	0,98966	0,97518	FCTH	1,00000	1,00000	1,00000	0,99351	0,98338

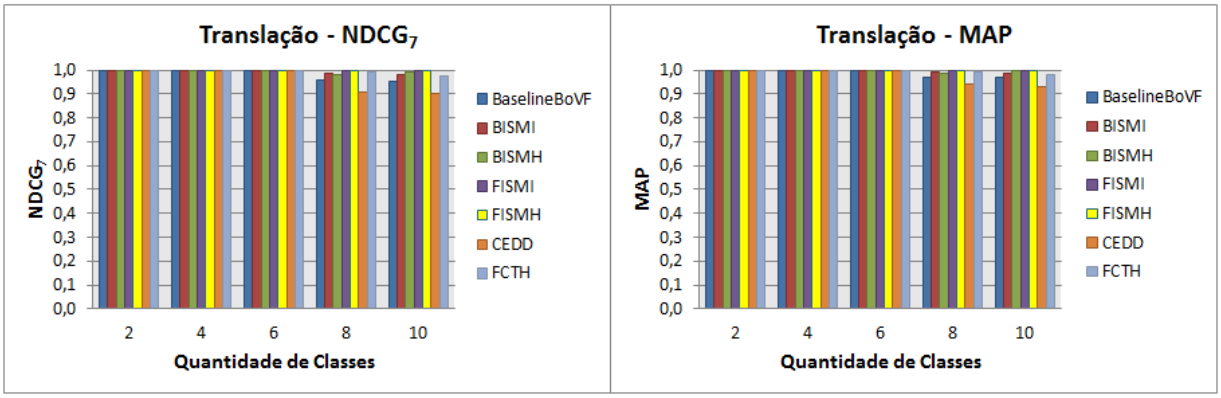

Figura 5.19: Resultados para a base de Translação variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 7$ considerando o total de imagens relevantes deste experimento.

FUNDO - $NDCG_{11}$						FUNDO - MAP					
Método	Quantidade de Classes					Método	Quantidade de Classes				
	2	4	6	8	10		2	4	6	8	10
BaselineBoVF	0,71214	0,76428	0,80838	0,76680	0,74246	BaselineBoVF	0,76341	0,78662	0,81264	0,77201	0,74223
BISMI	0,81473	0,88824	0,92249	0,91500	0,89025	BISMI	0,84481	0,90018	0,92009	0,91679	0,89527
BISMH	0,79167	0,87613	0,91649	0,91077	0,88746	BISMH	0,81876	0,89429	0,91288	0,91366	0,89395
FISMI	0,82491	0,88352	0,92091	0,90890	0,89114	FISMI	0,84446	0,88711	0,91482	0,91490	0,89342
FISMH	0,84758	0,91562	0,94481	0,94378	0,92204	FISMH	0,86029	0,91978	0,94329	0,93871	0,92058
CEDD	0,75916	0,56756	0,38631	0,33156	0,30676	CEDD	0,80277	0,57346	0,40620	0,34230	0,31097
FCTH	0,71501	0,51837	0,37905	0,31793	0,29420	FCTH	0,75511	0,52367	0,37985	0,32170	0,28820

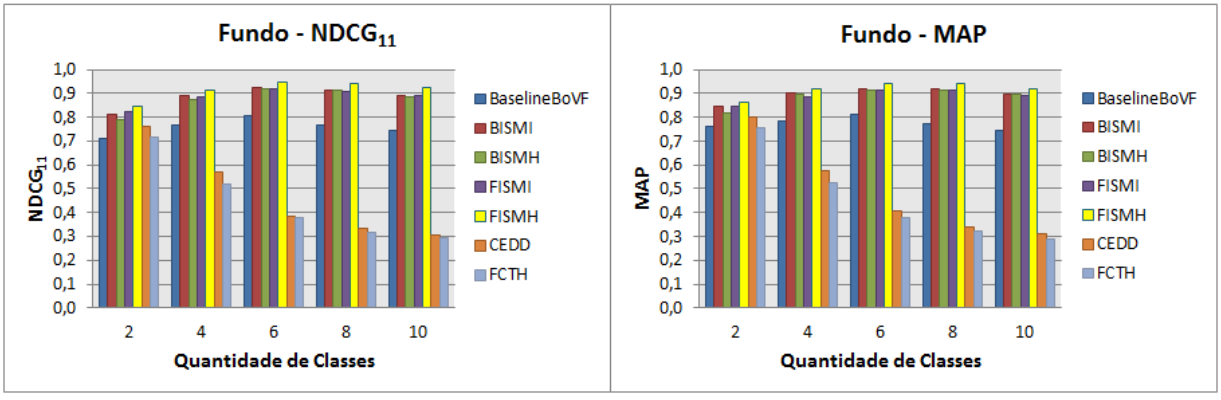

Figura 5.20: Resultados para a base de Fundo variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 11$ considerando o total de imagens relevantes deste experimento.

HÍBRIDO - $NDCG_{155}$					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,89605	0,89256	0,89140	0,82797	0,77056
BISMI	0,93192	0,93895	0,94776	0,91419	0,88594
BISMH	0,94134	0,94138	0,94850	0,91031	0,87929
FISMI	0,93676	0,93313	0,94302	0,91464	0,88589
FISMH	0,93912	0,95084	0,95091	0,92685	0,89737
CEDD	0,84967	0,74169	0,65708	0,60369	0,58637
FCTH	0,86770	0,73611	0,66515	0,61865	0,59791

HÍBRIDO - MAP					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,94037	0,91416	0,91531	0,84521	0,78615
BISMI	0,96433	0,95094	0,95447	0,92377	0,89826
BISMH	0,96785	0,95462	0,95495	0,92122	0,89276
FISMI	0,96524	0,94952	0,95090	0,92816	0,89771
FISMH	0,96759	0,95757	0,95729	0,93685	0,90934
CEDD	0,89577	0,77458	0,65491	0,56216	0,53028
FCTH	0,91558	0,77646	0,67684	0,58696	0,54514

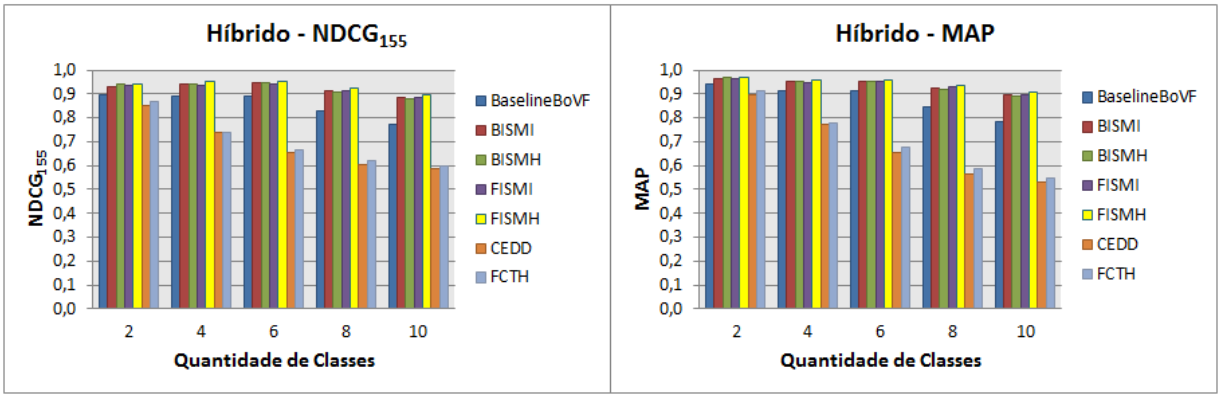

Figura 5.21: Resultados para a base de Híbrido variando a quantidade de classes utilizadas. Para a medida $NDCG_k$, $k = 155$ considerando o total de imagens relevantes deste experimento.

CENA - MAP					
Método	Quantidade de Classes				
	2	4	6	8	10
BaselineBoVF	0,93220	0,88387	0,87114	0,8634	0,85392
BISMI	0,92485	0,88604	0,89835	0,9071	0,89196
BISMH	0,92686	0,89054	0,87956	0,8954	0,88575
FISMI	0,93136	0,88393	0,88031	0,8725	0,86679
FISMH	0,93076	0,88920	0,88533	0,8762	0,87122
CEDD	0,87271	0,75752	0,60521	0,45948	0,40933
FCTH	0,88820	0,75569	0,60974	0,46812	0,40850

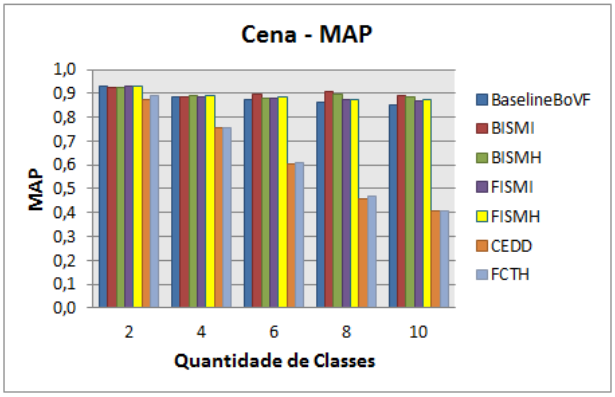

Figura 5.22: Resultados para a base de Cena variando a quantidade de classes utilizadas.

5.4 Considerações Finais

Neste capítulo foram apresentados nove experimentos com o intuito de avaliar o desempenho dos métodos BISMI, FISMI, BISMH e FISMH propostos neste trabalho.

Os experimentos dos métodos propostos na base de dados Wang, apesar de obterem resultados superiores a proposta BaselineBoVF considerada como *baseline*, não tiveram resultados satisfatórios. Para este tipo de base, que não possui uma boa representação na abordagem *bag-of-visual-features*, os métodos propostos não são bons, no entanto apresentam resultados bastante promissores para bases de versões. A recuperação de versões tem sido de interesse da polícia federal principalmente para identificar violação de *copyright*, ou adulteração de imagens.

Avaliando os experimento das bases de versões utilizando como medidas de avaliação o MAP e o $NDCG_k$, constatamos que a proposta de se utilizar as técnicas de extração da percepção visual humana para ajudar na recuperação de imagem por conteúdo tanto na abordagem binária (BISMI e BISMH) quanto na abordagem *fuzzy* (FISMI e FISMH) são eficientes e se mostram pouco variantes a rotação, translação, escala, cor, iluminação e mudança de fundo. Foi também mostrado a capacidade de flexibilizar a busca do usuário de acordo com a sua necessidade permitindo-o realizar consultas onde o interesse seja o *foreground* da imagem ao invés do *background* e vice-versa.

Portanto, avaliando os resultados dos experimentos realizados nessas oito bases de versões apresentados neste Capítulo, os métodos BISMI, FISMI, BISMH e FISMH mostram um bom desempenho no processo de recuperação de imagem por conteúdo superando a estratégia BaselineBoVF e os descritores CEDD e FCTH.

Capítulo 6

Conclusão e Trabalhos Futuros

Este capítulo apresenta as conclusões obtidas com base nos diversos experimentos realizados ao longo deste trabalho. No final do capítulo são apresentadas possíveis linhas de pesquisas futuras com o intuito de se evoluir a proposta apresentada.

6.1 Conclusão

Neste trabalho propusemos duas estratégias diferentes para construir o histograma de frequência de palavras visuais utilizando a abordagem *bag-of-visual-features*. Nessas estratégias utilizam-se as informações dos mapas de saliência das imagens, conforme proposto por [Itti et al. 1998] e [Harel et al. 2007], com o intuito de separar e ponderar a importância do *foreground* e *background* da imagem. No processo de separação e ponderação utilizaram-se duas estratégias, a binária na qual se estabelece um *trheshold* no mapa de saliência para separar as palavras visuais que ocorrem no *background* e *foreground* da imagem e a estratégia *fuzzy* na qual o grau de importância da palavra visual na imagem é definido pelo mapa de saliência, separando de forma nebulosa o *background* do *foreground*. Foi visto que essas estratégias apesar de não apresentarem bons resultados na base de dados Wang, foram eficientes nas mais diversas bases de versões de imagens com diferentes condições de iluminação, cor, escala, rotação, translação e mudança de fundo melhorando o poder discriminativo dos histogramas gerados.

Outra contribuição deste trabalho foi a criação de bases de versões de imagens não encontradas na literatura e necessária para avaliar as estratégias propostas nos diversos contextos. Ao todo foram criadas oito bases de imagens (Iluminação, Cor, Escala, Rotação, Translação, Fundo, Híbrido e Cena) com um total de 4720 imagens e 10 classes diferentes.

Também foi possível avaliar a capacidade e a flexibilidade das propostas apresentadas em possibilitar o usuário escolher o foco de sua pesquisa por similaridade podendo ponderar o seu interesse no *background* ou *foreground* das imagens a serem analisadas mantendo o bom desempenho da recuperação.

6.2 Trabalhos Futuros

Com relação as medidas de avaliação utilizadas, os métodos propostos neste trabalho mostraram superioridade em relação a outros métodos na literatura. No entanto, as propostas aqui apresentadas podem ser evoluídas com pesquisas futuras.

Neste trabalho focou-se no processo de construção de histogramas de frequências de palavras visuais da abordagem *bag-of-visual-features*, porém muitas melhorias ainda podem ser obtidas no processo de construção do dicionário de palavras visuais. A sugestão seria explorar maneiras diferentes de se agrupar o conjunto de características para gerar um dicionário mais discriminativo. Uma das possíveis maneiras seria a exploração dos Mapas Auto-Organizáveis de Kohonen [Kohonen 1990] como mostrado no trabalho de Kinnunen [Kinnunen et al. 2009].

Outra possível linha de pesquisa seria a melhoria e/ou utilização de outras técnicas de extração de mapas de saliência das imagens pois neste trabalho foram utilizadas somente as técnicas de [Itti et al. 1998] e [Harel et al. 2007]. Além disso, poderia ser explorado também a combinação de diversas técnicas de extração dos mapas de saliência podendo, por exemplo, testar a combinação por união ou intersecção de mapas de saliência extraídos por estratégias diferentes.

Referências Bibliográficas

[Abbasi et al. 2000] Abbasi, S., Mokhtarian, F., e Kittler, J. (2000). Enhancing CSS-based shape retrieval for objects with shallow concavities. *Image and Vision Computing*, 18(3):199–211.

[Ahmed et al. 1974] Ahmed, N., Natarajan, T., e Rao, K. R. (1974). Discrete Cosine Transfom. *IEEE Transactions on Computers*, 23(1):90–93.

[Almeida et al. 2009] Almeida, J., Torres, R., e Goldenstein, S. (2009). SIFT applied to CBIR. *Revista de Sist. de Inf. da Fac. Salesiana Maria Auxiliadora*, 4:41-48.

[Alto et al. 2005] Alto, H., Rangayyan, R. M., e Desautels, J. E. L. (2005). Content-based retrieval and analysis of mammographic masses. *Journal of Electronic Imaging*, 14(2):023016.

[Arivazhagan e Ganesan 2003] Arivazhagan, S. e Ganesan, L. (2003). Texture classification using wavelet transform. *Pattern Recognition Letters*, 24(9-10):1513–1521.

[Avidan e Shamir 2007] Avidan, S. e Shamir, A. (2007). Seam carving for content-aware image resizing. In *ACM Transactions on Graphics*, p. 10. SIGGRAPH.

[Baeza-Yates e Ribeiro-Neto 1999] Baeza-Yates, R. A. e Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Batista et al. 2009] Batista, N. C., Lopes, A. P. B., e de A. Araujo, A. (2009). Detecting Buildings in Historical Photographs Using Bag-of-Keypoints. *Computer Graphics and Image Processing, Brazilian Symposium on*, 0:276–283.

[Bay et al. 2006] Bay, H., Tuytelaars, T., e Gool, L. V. (2006). SURF: Speeded up robust features. In *In European Conference on Computer Vision, ECCV*, pp. 404–417.

[Bober 2001] Bober, M. (2001). MPEG-7 Visual Shape Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):716–719.

[Borg e Groenen 2005] Borg, I. e Groenen, P. (2005). *Modern Multidimensional Scaling: Theory and Applications*. Springer Series in Statistics. Springer.

[Brodatz 2012] Brodatz (2012). Brodatz textures. `http://www.ux.uis.no/~tranden/brodatz.html` [Online. Último acesso: 2012-12-16].

[Canny 1986] Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

[Carson et al. 1999] Carson, C., Thomas, M., Belongie, S., Hellerstein, J. M., e Malik, J. (1999). Blobworld: A System for Region-Based Image Indexing and Retrieval. In *Third International Conference on Visual Information Systems*, pp. 509–516. Springer.

[Chang e Fu 1980] Chang, N.-S. e Fu, K.-S. (1980). Query-by-Pictorial-Example. *IEEE Transactions on Software Engineering*, 6(6):519–524.

[Chang e Kunil 1981] Chang, S.-K. e Kunil, T. L. (1981). Pictorial Data-Base Systems. *Computer*, 14(11):13–21.

[Chatzichristofis e Boutalis 2007] Chatzichristofis, S. e Boutalis, Y. (2007). A hybrid scheme for fast and accurate image retrieval based on color descriptors. In *Proceedings of The Eleventh IASTED International Conference on Artificial Intelligence and Soft Computing*, pp. 280–285, Palma de Mallorca, Spain, Anaheim, CA, USA. ACTA Press.

[Chatzichristofis et al. 2010] Chatzichristofis, S. A., Zagoris, K., Boutalis, Y. S., e Papamarkos, N. (2010). Accurate Image Retrieval Based on Compact Composite Descriptors and Relevance Feedback Information. *International Journal of Pattern Recognition and Artificial Intelligence, IJPRAI*, 24(2):207–244.

[Csurka et al. 2004] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., e Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, European Conference on Computer Vision, ECCV*, pp. 1–22.

[Daubechies 1990] Daubechies, I. (1990). The Wavelet Transform, Time-Frequency Localisation and Signal Analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005.

[de Berg et al. 2008] de Berg, M., Cheong, O., van Kreveld, M., e Overmars, M. (2008). *Computational Geometry Algorithms and Applications*. Springer-Verlag, Santa Clara, CA, USA, 3 edition.

[Dumais et al. 1998] Dumais, S., Platt, J., Sahami, M., e Heckerman, D. (1998). Inductive Learning Algorithms and Representations for Text Categorization. In *Proceedings of the seventh international Conference on Information and Knowledge Management, CIKM*, pp. 148–155, Bethesda, Maryland, United States. ACM Press.

[Everingham et al. 2005] Everingham, M., Zisserman, A., Williams, C., Gool, L. V., Allan, M., Bishop, C., Chapelle, O., Dalal, N., Deselaers, T., Dorko, G., et al. (2005). The 2005 PASCAL visual object classes challenge. *First PASCAL Challenge Workshop*.

[Fayyad e Uthurusamy 2002] Fayyad, U. e Uthurusamy, R. (2002). Evolving data into mining solutions for insights. *Communications of the ACM*, 45:28–31.

[Fei-Fei et al. 2007] Fei-Fei, L., Fergus, R., e Perona, P. (2007). Learning Generative Visual Models from Few Training Examples: An Incremental Bayesian Approach Tested on 101 Object Categories. *Computer Vision and Image Understanding*, 106(1):59–70.

[Feng D. 2003] Feng D., Siu W. C., Z. H. (2003). *Multimedia Information Retrieval and Management: Technological Fundamentals and Applications*. Springe-Verlag Berlin Heidelberg New York.

[Fischer e Weber 1993] Fischer, B. e Weber, H. (1993). Express Saccades and Visual-Attention. *Behavioral and Brain Sciences*, 16(3):553–567.

[Gonzalez e Woods 2006] Gonzalez, R. C. e Woods, R. E. (2006). *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

[Greenspan et al. 1994] Greenspan, H., Belongie, S., Goodman, R., Perona, P., Rakshit, S., e Anderson, C. H. (1994). Overcomplete steerable pyramid filters and rotation invariance. In *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 222–228.

[Griffin et al. 2007] Griffin, G., Holub, A., e Perona, P. (2007). Caltech-256 Object Category Dataset. Technical Report 7694, California Institute of Technology.

[Harel et al. 2007] Harel, J., Koch, C., e Perona, P. (2007). Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pp. 545–552. MIT Press.

[Huang et al. 2008] Huang, J.-H., Zia, A., Zhou, J., e Robles-Kelly, A. (2008). Content-Based Image Retrieval via Subspace-Projected Salient Features. pp. 593–599, Washington, DC, USA.

[Itti e Koch 2001] Itti, L. e Koch, C. (2001). Computational Modelling of Visual Attention. *Nature Reviews Neuroscience*, 2(3):194–203.

[Itti et al. 1998] Itti, L., Koch, C., e Niebur, E. (1998). A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259.

[Jacobs et al. 2000] Jacobs, D. W., Weinshall, D., e Gdalyahu, Y. (2000). Classification with Nonmetric Distances: Image Retrieval and Class Representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(6):583–600.

[Jain et al. 1999] Jain, A. K., Murty, M. N., e Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323.

[Järvelin e Kekäläinen 2000] Järvelin, K. e Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '00, pp. 41–48, Athens, Greece, New York, NY, USA. ACM.

[Jégou et al. 2010] Jégou, H., Douze, M., e Schmid, C. (2010). Improving Bag-of-Features for Large Scale Image Search. *International Journal of Computer Vision*, 87(3):316–336.

[Jurie e Schmid 2004] Jurie, F. e Schmid, C. (2004). Scale-invariant shape features for recognition of object categories. In *International Conference on Computer Vision & Pattern Recognition*, volume II, pp. 90–96.

[Kato 1992] Kato, T. (1992). Database architecture for content-based image retrieval. *Image Storage and Retrieval Systems Proc SPIE*, 1662:112–123.

[Ke e Sukthankar 2004] Ke, Y. e Sukthankar, R. (2004). PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 2:506–513.

[Kimura et al. 2011] Kimura, P. A. S., Cavalcanti, J. M. B., Saraiva, P. C., da Silva Torres, R., e Gonçalves, M. A. (2011). Evaluating Retrieval Effectiveness of Descriptors for Searching in Large Image Databases. *Journal of Information and Data Management, JIDM*, 2(3):305–320.

[Kinnunen et al. 2009] Kinnunen, T., Kamarainen, J.-K., Lensu, L., e Kälviäinen, H. (2009). Bag-of-Features Codebook Generation by Self-Organisation. In *Proceedings of the 7th International Workshop on Advances in Self-Organizing Maps*, WSOM '09, pp. 124–132, St. Augustine, FL, USA, Berlin, Heidelberg. Springer-Verlag.

[Koch e Ullman 1985] Koch, C. e Ullman, S. (1985). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. In *Human Neurobiology*, volume 4, pp. 219–227.

[Kohonen 1990] Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480.

[Kotoulas e Andreadis 2005] Kotoulas, L. e Andreadis, I. (2005). Image analysis using moments. *Information Systems Journal*, 1(1):360–364.

[Kutz et al. 2003] Kutz, O., Sturm, H., Suzuki, N.-Y., Wolter, F., e Zakharyaschev, M. (2003). Logics of metric spaces. *ACM Transactions on Computational Logic (TOCL)*, 4(2):260–294.

[Lazebnik et al. 2003] Lazebnik, S., Schmid, C., e Ponce, J. (2003). Affine-Invariant Local Descriptors and Neighborhood Statistics for Texture Recognition. In *In Proc. ICCV*, pp. 649–655.

[Liang et al. 2010] Liang, Z., Fu, H., Chi, Z., e Feng, D. D. (2010). Salient-SIFT for Image Retrieval. In *Advanced Concepts for Intelligent Vision Systems, ACIVS*, pp. 62–71. Springer.

[Lindeberg 1993] Lindeberg, T. (1993). Detecting Salient Blob-Like Image Structures and Their Scales with a Scale-Space Primal Sketch: A Method for Focus-of-Attention. *International Journal of Computer Vision*, 11:283–318.

[Liu et al. 2007] Liu, Y., Zhang, D., Lu, G., e Ma, W. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282.

[Lopes et al. 2009a] Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., e de Albuquerque Araújo, A. (2009a). A bag-of-features approach based on Hue-SIFT descriptor for nude detection. In *EUSIPCO*, pp. 1552–1556. EUSIPCO European Signal Processing Conference.

[Lopes et al. 2009b] Lopes, A. P. B., de Avila, S. E. F., Peixoto, A. N. A., Oliveira, R. S., de M. Coelho, M., e de Albuquerque Araújo, A. (2009b). Nude Detection in Video Using Bag-of-Visual-Features. In *SIBGRAPI*, pp. 224–231. IEEE Computer Society.

[Lowe 2004] Lowe, D. G. (2004). Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110.

[Lux e Chatzichristofis 2008] Lux, M. e Chatzichristofis, S. A. (2008). Lire: lucene image retrieval: an extensible java CBIR library. In *Proceedings of the 16th ACM international conference on Multimedia*, MM '08, pp. 1085–1088, Vancouver, British Columbia, Canada, New York, NY, USA. ACM.

[Manjunath et al. 2001] Manjunath, B. S., rainer Ohm, J., Vasudevan, V. V., e Yamada, A. (2001). Color and Texture Descriptors. *IEEE Transactions on Circuits and Systems for Video Technology*, 11(6):703–715.

[Manning et al. 2008] Manning, C. D., Raghava, P., e Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

[Marée et al. 2005] Marée, R., Geurts, P., Piater, J., e Wehenkel, L. (2005). Random subwindows for robust image classification. In *Computer Vision and Pattern Recognition, CVPR*, pp. 34–40. IEEE.

[Marszalek e Schmid 2006] Marszalek, M. e Schmid, C. (2006). Spatial Weighting for Bag-of-Features. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, CVPR 06, pp. 2118–2125, Washington, DC, USA. IEEE Computer Society.

[Moosmann et al. 2006] Moosmann, F., Larlus, D., e Jurie, F. (2006). Learning saliency maps for object categorization. In *ECCV International Workshop on The Representation and Use of Prior Knowledge in Vision*. Springer.

[Nakamoto e Toriu 2011] Nakamoto, S. e Toriu, T. (2011). Combination Way of Local Properties, Classifiers and Saliency in Bag-of-Keypoints Approach for Generic Object Recognition. *International Journal of Computer Science and Network Security*, 11:35–42.

[Nascimento et al. 2003] Nascimento, M. A., Sridhar, V., e Li, X. (2003). Effective and efficient region-based image retrieval. *Journal of Visual Languages and Computing*, 14:151–179.

[Niblack et al. 1993] Niblack, C. W., Barber, R. J., Equitz, W. R., Flickner, M. D., Glasman, D., Petkovic, D., e Yanker, P. C. (1993). The QBIC Project: Querying Image by Content Using Color, Texture, and Shape. 1908(1):173–187.

[Nilsback e Zisserman 2006] Nilsback, M.-E. e Zisserman, A. (2006). A Visual Vocabulary for Flower Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pp. 1447–1454.

[Nister e Stewenius 2006] Nister, D. e Stewenius, H. (2006). Scalable Recognition with a Vocabulary Tree. *Computer Vision and Pattern Recognition*, 2:2161–2168.

[Nowak et al. 2006] Nowak, E., Jurie, F., e Triggs, B. (2006). Sampling strategies for bag-of-features image classification. In *European Conference on Computer Vision*, volume 3954, pp. 490–503, Graz, Austria.

[Opelt e Pinz 2005] Opelt, A. e Pinz, A. (2005). Object localization with boosting and weak supervision for generic object recognition. In *Proceedings of the 14th Scandinavian Conference on Image Analysis (SCIA)*, pp. 862–871.

[Otsu 1979] Otsu, N. (1979). A Threshold Selection Method from Gray-Level Histograms. *Systems, Man and Cybernetics, IEEE Transactions on*, 9(1):62–66.

[Pass et al. 1996] Pass, G., Zabih, R., e Miller, J. (1996). Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*, MULTIMEDIA '96, pp. 65–73, Boston, Massachusetts, United States, New York, NY, USA. ACM.

[Penatti e da Silva Torres 2008] Penatti, O. A. B. e da Silva Torres, R. (2008). Color Descriptors for Web Image Retrieval: A Comparative Study. In *Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI*, pp. 163–170.

[Ribeiro-Neto et al. 1999] Ribeiro-Neto, B., Moura, E. S., Neubert, M. S., e Ziviani, N. (1999). Efficient distributed algorithms to build inverted files. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '99, pp. 105–112, Berkeley, California, United States, New York, NY, USA. ACM.

[Rohani e Nugroho 2008] Rohani, B. e Nugroho, B. (2008). Manhattan-Chebychev Distance Metric for MIMO Systems. In *IEICE Technical Report*, volume 108, pp. 49–52, Ishikawa.

[Rui et al. 1997] Rui, Y., Huang, T. S., e Chang, S.-f. (1997). Image Retrieval: Past, Present, And Future. In *Journal of Visual Communication and Image Representation*, volume 10, pp. 1–23.

[Santini e Gupta 2001] Santini, S. e Gupta, A. (2001). A Wavelet Data Model For Image Databases. *Multimedia and Expo, IEEE International Conference on*, p. 280.

[Santini e Jain 1996] Santini, S. e Jain, R. (1996). Gabor Space and the Developement of Preattentive Similarity. In *International Conference on Pattern Recognition, ICPR*, pp. 40–, Washington, DC, USA. IEEE Computer Society.

[Sato e Katto 2010] Sato, M. e Katto, J. (2010). Performance improvement of generic object recognition by using seam carving and saliency map. In *The International Workshop on Advanced Image Technology, IWAIT*.

[Shilane et al. 2004] Shilane, P., Min, P., Kazhdan, M., e Funkhouser, T. (2004). The Princeton Shape Benchmark. In *In Shape Modeling International*, pp. 167–178.

[Sivic e Zisserman 2003] Sivic, J. e Zisserman, A. (2003). Video Google: A Text Retrieval Approach to Object Matching in Videos. *Computer Vision, IEEE International Conference on*, 2:1470.

[Smith e fu Chang 1996] Smith, J. R. e fu Chang, S. (1996). Tools and Techniques for Color Image Retrieval. In *Conference on Storage and Retrieval for Image and Video Database IV, SPIE*, pp. 426–437.

[Sonka et al. 1998] Sonka, M., Hlavac, V., e Boyle, R. (1998). *Image Processing: Analysis and Machine Vision*. CL-Engineering, 2 edition.

[Stehling et al. 2002] Stehling, R. O., Nascimento, M. A., e Falcão, A. X. (2002). A compact and efficient image retrieval approach based on border/interior pixel classification. In *Proceedings of the eleventh international conference on Information and knowledge management*, CIKM '02, pp. 102–109, McLean, Virginia, USA, New York, NY, USA. ACM.

[Stollnitz et al. 1996] Stollnitz, E. J., Derose, T. D., e Salesin, D. H. (1996). *Wavelets for computer graphics: theory and applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Swain e Ballard 1991] Swain, M. J. e Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, 7(1):11–32.

[Traina et al. 2003] Traina, A. J. M., Traina, C., Bueno, J. M., Chino, F. J. T., e Azevedo-Marques, P. (2003). Efficient Content-Based Image Retrieval through Metric Histograms. *World Wide Web*, 6(2):157–185.

[Wang et al. 2001] Wang, J. Z., Li, J., e Wiederhold, G. (2001). SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(9):947–963.

[Wang et al. 2008] Wang, L., Zhou, L., e Shen, C. (2008). A Fast Algorithm for Creating a Compact and Discriminative Visual Codebook. In *European Conference on Computer Vision (ECCV'08)*, volume 4, pp. 719–732, Marseille, France. Lecture Notes in Computer Science (LNCS) 5305.

[Winn et al. 2005] Winn, J., Criminisi, A., e Minka, T. (2005). Object Categorization by Learned Universal Visual Dictionary. In *Proceedings of the Tenth IEEE International Conference on Computer Vision - Volume 2*, ICCV '05, pp. 1800–1807, Washington, DC, USA.

[Yang et al. 2007] Yang, J., Jiang, Y.-G., Hauptmann, A. G., e Ngo, C.-W. (2007). Evaluating bag-of-visual-words representations in scene classification. In Wang, J. Z., Boujemaa, N., Bimbo, A. D., e Li, J. (editores), *Multimedia Information Retrieval*, pp. 197–206.

[Yang et al. 2009] Yang, X., Zhu, Q., e Cheng, K.-T. (2009). MyFinder: near-duplicate detection for large image collections. In Gao, W., Rui, Y., Hanjalic, A., Xu, C., Steinbach, E. G., El-Saddik, A., e Zhou, M. X. (editores), *ACM Multimedia*, pp. 1013–1014.

[Zezula et al. 2006] Zezula, P., Amato, G., Dohnal, V., e Batko, M. (2006). *Similarity Search: The Metric Space Approach*, volume 32 de *Advances in Database Systems*. Springer.

[Zhang e Chang 2004] Zhang, D.-Q. e Chang, S.-F. (2004). Detecting image near-duplicate by stochastic attributed relational graph matching with learning. In *Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 877–884, New York, NY, USA, New York, NY, USA. ACM.